Informatic analysis of proteins with a role in oxidative damage and ageing

Tina Ann Eyre

Department of Biology
University College London

A thesis submitted to the University of London in the Faculty of Science for the degree of Doctor of Philosophy

October 2004

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

Ageing is a complex, universal process that remains very poorly understood, particularly in mammals. This thesis attempts to increase our understanding of ageing by predicting the structure of the uncoupling proteins, membrane proteins with a possible role in the modulation of oxidative damage, and therefore of ageing. A 3-dimensional model of the uncoupling proteins is generated, based on an analysis of known membrane proteins structures. In order to assess the accuracy of this model it is compared to the actual structure of a homologous protein, solved after the modelling was complete. A homology packing model is produced that, in combination with predictions of likely functional residues, will be of use in establishing the mechanism of action of the uncoupling proteins.

Additionally, this thesis investigates the regulation of ageing by the insulin-like signalling pathway and the transcription factor DAF-16. Longevity- and ageing-associated transcription factor binding sites are identified, due to their over- or under-representation within genes regulated by this pathway. Direct and indirect DAF-16 target gene classes are identified, and possible mechanisms of feedback control of the pathway are investigated, including the identification of other transcription factors whose expression is regulated by DAF-16. Although this work is a valuable starting point, considerable further work will be required before a full understanding of the regulation of ageing is obtained.

Finally, this thesis has provided insights into membrane protein structure and its prediction. A comprehensive analysis of these structures was performed and the results used to develop a modelling method that is applicable to structure prediction for all membrane proteins. Although buried transmembrane helix faces were identified with relatively high accuracy, a greater understanding of membrane protein structure is required before reliable 3-dimensional models can be produced using this method. The opening of a structural genomics project focused on membrane proteins is helping to bring the realisation of this aim closer to the present.

This work was generously supported by the Biotechnology and Biological Sciences Research Council.

Acknowledgements

UCL has been a wonderful place to work and this has been entirely thanks to the people who have been around to help, encourage, advise (and lead me astray) throughout my PhD.

Many thanks go to everyone who has been involved, but especially to my supervisors Janet and Linda, without whom none of it would have been possible. Both Linda and Janet have been there with advice, support and inspiration whenever it was needed, but have also allowed me the freedom in my work that I have so enjoyed.

Thanks also all of those who have given up their time to help me out and have made UCL such a fun place to work. These have included all members of Janet's and Christine Orengo's groups, both past and present, in London and in Hinxton, including Al Grant, Adrian EkoUkeh, Andrew Harrison, Annabel Todd, Brian Ferguson, Chris Bennett, Christine Orengo, Dan Buchan, Dave Lee, Duncan McKenzie, Donovan Binns, Eric Blanc, Eugene Schuster, Frances Pearl, Gabby Reeves, Gail Bartlet, Gareth Stockwell, Gordon Whammond, Hannes Ponstingl, Hugh Shanahan, Ian Sillitoe, Ilhem Diboun, Jahid Ahmed, James Bray, Jen Dawe, Jessie Oldershaw, Juan Antonio, Mark Dibley, Ollie Redfern, Roman Laskowski, Russell Marsden, Sarah Addou, Stathis Sidero, Stefano Lise, Stuart Rison, Sue Jones, Thomas Kabir and Tim Dallman.

Particular thanks to Ollie for cheering me up and amusing me with tales of his antics, and for all of his help. Also thanks to Gabby, Russell, Stuart and James for passing on the BSM LaTeX secrets to the next generation, reading draft chapters and for all of the other help, which I couldn't have done without. For advice on maths, use of data and programs and helpful discussion, thanks to Roman, Hugh, Hannes, Harry, Sue, Thomas and Josh McElwee. Finally in terms of work, thanks Gene for all of your help with Chapter 6; I couldn't have done it without you.

Moving on to the social side of life at UCL, thanks to Christine for adopting me into her group, to all GB drinkers (sorry if the new generation strayed a little in search of our beloved ribs!) and all of those who made life fun, including, but not limited to, Kim, Ro, Mariana, Ollie, Dave, Stuart, Gabby, James and Dan.

Special thanks to my parents for giving me the love and encouragement to enable me to begin a PhD. And finally, the biggest thanks go to my husband, Roger, who has enabled me to finish it. Without your help, support, and ability to always make me smile, I could never have done it.

Love and thanks to everyone.

Tina Eyre, October 2004.

Contents

List of Figures

List of Tables

List of abreviations used

Abbreviation	Definition
3D	3-dimensional
ADP	Adenosine diphosphate
ANT	Adenine nucleotide translocase
ASA	Accessible surface area
ATP	Adenosine triphosphate
BMCP1	Brain mitochondrial carrier protein 1
bp	Base pair
CNS	Central nervous system
CYP	Cytochrome P450
DAE	DAF-16-associated element
DBE	DAF-16 binding element
DNA	Deoxyribrose nucleic acid
DUF	Domain of unknown function
GES scale	Goldman Engelman Scale
GR	Glucocorticoid receptor
GST	Glutathione-S-transferase
HRE	HIF-1 response element
HSAS	Heat shock-associated sequence
HSE	Heat shock element
HSF	Heat shock factor
IGF	Insulin-like growth factor
ILS	Insulin/IGF-like signalling
IMM	Inner mitochondrial membrane
IMS	Inter-membrane space
KD scale	Kyte and Doolittle hydrophobicity scale
kDa	kDaltons
LA scale	Lipid-tail-accessibility scale
MCF	Mitochondrial carrier protein family
NDP	Nucleotide diphosphate
NTP	Nucleotide triphosphate
P	Probability
PDB	Protein data bank
PMF	Proton motive force

Table 2.1: *continued*

Abbreviation	Definition
PPAR	Peroxisome proliferator activated receptor
RNAi	Ribonucleic acid interference
ROS	Reactive oxygen species
SOD	Superoxide dismutase
TF	Transcription factor
TFBS	Transcription factor binding site
TM	Transmembrane
UCP	Uncoupling protein
UGT	UDP-glucuronosyl transferase
WW scale	White and Wimley hydrophobicity scale

List of amino acid abreviations

A	Ala	Alanine
C	Cys	Cysteine
D	Asp	Aspartate
E	Glu	Glutamate
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
Y	Tyr	Tyrosine

Declaration

All research presented in this thesis is the candidate's own work. The content of Chapter 3 was previously published as:

Eyre, T.A., Partridge, L. and Thornton, J.M. (2004) Analysis of alpha-helical transmembrane protein structure: Implications for prediciton of 3D structural models. *Protein Eng Des Sel.*, **17**, 613-24.

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of University College London or any other university or institute of learning.

Chapter 1

Introduction

1.1 Ageing

1.1.1 What is ageing?

Ageing affects us all. It is a phenomenon universal throughout the world and probably across all of the animal kingdoms of life. It has a huge social and economic impact, in terms of National Health Service and research spending and loss of the workforce through retirement and age-related disease. Despite this, ageing has been defined in many different ways. For example, ageing has been defined as a decline in the efficiency of physiological processes after the reproductive phase of life (Halliwell & Gutteridge, 1999). Alternatively, ageing is said to describe the decline in survival and fecundity (fertility) with advancing age (Partridge & Gems, 2002). Another definition is that ageing consists of a set of early-onset, slowly progressive, mutually synergistic degenerative processes (de Grey *et al.*, 2002). Finally, ageing can be thought of more simply as the series of processes that ultimately end life (Busuttil *et al.*, 2004).

What all of these definitions have in common is that ageing is associated with deterioration. Ageing is a process that has evolved across a wide variety of species. There has therefore been much discussion as to why such a phenomenon would evolve it has only a detrimental effect on the individual. Several theories have been put forward and, as described below, it seems likely that ageing arises as a result of an evolutionary trade off (Williams, 1957; Luckinbil *et al.*, 1984; Zwaan *et al.*, 1995).

The loss of fitness associated with ageing is often proposed to be the result of accumulation of damage throughout life. What form this damage takes, or what causes it, is still open to discussion, and it is possible that a great many agents and targets are involved. For example, many current theories favour the well established idea that ageing is related to oxidative damage caused by reactive oxygen species (ROS) (Harman & Piette, 1966).

This may lead to damage of DNA, proteins and lipids and impairment of cellular function. The major source of the ROS is thought to be the mitochondria. However, recent evidence has suggested that other factors may contribute to ageing, such as damage caused by xenobiotic compounds (McElwee *et al.*, 2004).

1.1.2 Why does ageing occur?

Early gerontologists debated whether or not ageing is a genetically programmed process. Differences in rates of ageing between species suggest that a different rate of ageing has evolved for each species and is controlled genetically. In addition, mutational studies which cause extension of lifespan show that some genes can modify the rate of ageing (Kenyon *et al.*, 1993; Kimura *et al.*, 1997a; Hertweck *et al.*, 2004; Friedman & Johnson, 1988; Morris *et al.*, 1996; Tatar *et al.*, 2001; Clancy *et al.*, 2001; Bluher *et al.*, 2003; Holzenberger *et al.*, 2003). However, it seems unlikely that genes would have evolved purely to cause damage and lead to ageing. One likely explanation is that ageing is caused by the effects of mutations that are detrimental only late in life. Since the detrimental effects of these mutations generally appear only after reproduction, they are not removed from the population by natural selection. Hence, the rate of ageing is not genetically programmed but it will be affected by mutations in genes that affect the rate of damage or of repair.

The idea that ageing is the result of mutations with a detrimental effect late is life is often referred to as the Evolutionary Theory of ageing. Within this theory, two main branches exist, the Mutation-accumulation Theory (Hamilton, 1966) and the Trade-off or Pleiotropy Theory (Williams, 1957). The former refers to the accumulation to mutations whose only effect is a deleterious one, late in life (Hamilton, 1966). In contrast, the Pleiotropy Theory states that some mutations may have beneficial effects in youth but detrimental effects later. Since selection will act more strongly on the early than the late effect, these mutations will be maintained within the population (Williams, 1957).

One way to assess the relative likelihood of the mutation-accumulation and trade-off theories is to examine the degree of evolutionary conservation of the mechanisms of ageing. Due to the stochastic nature of mutations, the mutation-accumulation theory is more likely to lead to lineage-specific mechanisms, because the same mutations are unlikely to occur independently. In contrast, trade-offs between the early and late effects of mutations are likely to occur by similar mechanisms across lineages (Partridge & Gems, 2002).

The mechanisms of ageing appear to be highly conserved across large evolutionary distances (Guarente & Kenyon, 2000; Partridge & Gems, 2002), in support of the Pleiotropy

theory. This finding is extremely important to experimental studies of ageing since it implies that data from studies using lower model organisms such as *Caenorhabditis elegans*, which have many advantages to higher animal studies, will still have relevance to human ageing.

In general, experimental evidence also tends to more strongly support the Pleiotropy or Trade-off theory of ageing. The majority of such studies have been performed using lines of *Drosophila* selected for longevity or for shorter life (Luckinbil *et al.*, 1984; Zwaan *et al.*, 1995). In accordance with the trade-off theory, the reduced longevity of the rapidly ageing flies is associated with a beneficial effect early in life, when fecundity is increased relative to the long-lived lines. The reduction of lifespan in the flies selected for rapid ageing can be prevented by sterilisation, implying that the rapid ageing is a direct consequence of increased fecundity (Sgro & Partridge, 1999). Similarly, in *C. elegans*, reduced fecundity and increased lifespan are observed during dietary restriction (Zamiri, 1978; Chapman & Partridge, 1996; Chapman *et al.*, 1998; Sonntag *et al.*, 1999; Good & Tatar, 2001; Drummond-Barbosa & Spradling, 2001) and in long-lived strains of *Drosophila* (Luckinbil *et al.*, 1984; Zwaan *et al.*, 1995; Sgro & Partridge, 1999).

In summary, at present it is believed that ageing is the result of an evolutionary trade-off between genes that are beneficial early in life but detrimental late in life, or more specifically, a delayed cost of reproduction.

1.1.3 The Oxidative Damage Theory of ageing

The Oxidative Damage Theory of ageing was first proposed by Harman & Piette (1966). According to this now widely accepted theory, the damage that is the cause of ageing is the result of oxidative reactions between reactive oxygen species (ROS) and various cellular macromolecules. It was proposed that mitochondria are both the major source and target of such damage. The cell is protected from oxidative damage by antioxidant enzymes, such as superoxide dismutase (SOD) and catalase. These enzymes reduce levels of ROS by reaction of superoxide (O_2^-) with H^+ ions, producing water and oxygen:

SOD: $2O_2^- + 2H^+ \longrightarrow O_2 + H_2O_2$

Catalase: $2H_2O_2 \longrightarrow O_2 + 2H_2O$

Various strands of evidence provide support for the Oxidative Damage Theory of ageing. For example, several long-lived mutants in *C. elegans* (Murakami & Johnson, 1996; Honda & Honda, 1999) and in *Drosophila* (Hari *et al.*, 1998; Clancy *et al.*, 2001; Tatar *et al.*, 2001) show increased levels of SOD and catalase, and are more resistant to ROS and other stresses than wildtype animals.

To obtain further support for the Oxidative Damage Theory, antioxidant enzymes have been over-expressed and the consequent effects on lifespan determined. The results of these studies have often been contradictory, with several studies identifying little or no (Seto *et al.*, 1990; Orr & Sohal, 1993, 2003) increase in longevity, but others finding a marked 40-50% extension of lifespan (Orr & Sohal, 1994; Hari *et al.*, 1998; Parkes *et al.*, 1998a,b; Sun & Tower, 1999). Null mutations in SOD reduce longevity (Phillips *et al.*, 1989). In one study lifespan extension was achieved by over-expression specifically in the motorneurones of *Drosophila* (Parkes *et al.*, 1998a). Similarly contradictory results have been observed using catalase (Orr & Sohal, 1992).

The findings suggest that the roles of SOD and catalase in determining lifespan may be restricted to particular cell types or expression levels. The majority of studies have shown net oxidative damage to protein, lipids and DNA (Viteri *et al.*, 2004; Wozniak *et al.*, 2004; Gedik *et al.*, 2004) and levels of superoxide (Antier *et al.*, 2004) increase with age. Hence the role of oxidative damage in ageing, while well established, is far from being fully understood.

1.1.3.1 The uncoupling proteins

As described above, an increased ROS burden is believed to cause cellular damage that contributes to ageing. Where do these ROS derive from, and how is their production regulated? The uncoupling proteins (UCPs) are membrane proteins that have been proposed to form an essential part of the electron transport chain, modulating the production of superoxide, and consequently of ROS (Schrauwen *et al.*, 1999; Nishikawa *et al.*, 2000; Vidal-Puig *et al.*, 2000; Arsenijevic *et al.*, 2000; Mizuno *et al.*, 2000; Gong *et al.*, 2000; Casteilla *et al.*, 2001; Barazzoni & Nair, 2001; Kerner *et al.*, 2001). Some interesting studies have suggested that UCPs may play a role in the increased ROS damage that occurs during ageing, but the molecular mechanisms by which this occurs are unknown (Casteilla *et al.*, 2001).

UCPs are thought to catalyse leakage of protons through the inner mitochondrial membrane (IMM), back into the matrix, without adenosine triphosphate (ATP) synthesis. Hence the UCPs are said to 'uncouple' oxidative phosphorylation from ATP production. As uncoupling occurs, the energy stored in the proton gradient across the IMM is dissipated in the form of heat. This process forms the basis of non-shivering thermogenesis in a range of species, and is summarised in Figure 1.1.

Dissipation of the proton motive force (PMF) by the UCPs is thought to reduce the half-life of free radical intermediates of the electron transport chain. As a result, the likelihood that they will react with oxygen, producing ROS, is reduced and some of the

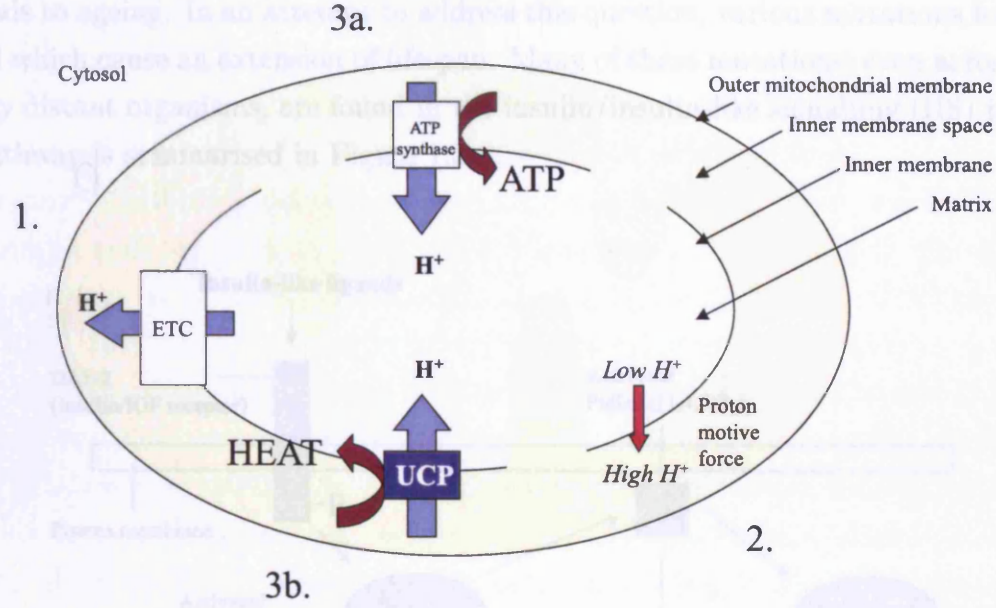


Figure 1.1: Uncoupling of ATP production by the UCPs. 1. Passage of electrons through the electron transport chain (ETC) leads to the pumping of protons out of the matrix. 2. This generates a gradient of proton concentration and charge across the inner mitochondrial membrane, known as a proton motive force (PMF). 3a. Energy stored in the PMF is used to drive the production of the energy-rich molecule ATP, as protons pass back down their gradient into the matrix through ATP synthase. 3b. In uncoupled mitochondria, some of the energy stored in the PMF is released as heat when protons leak back into the matrix via the UCPs, without the generation of ATP. This figure was drawn by hand using Microsoft PowerPoint.

damage that leads to ageing may be prevented (Casteilla *et al.*, 2001). The potential role of the UCPs in ageing requires further investigation. At the start of this work very little was known of either the structure or mechanism of action of these proteins. As a result of circular dichroism, computational topological and antigenic studies (Aquila *et al.*, 1985; Runswick *et al.*, 1987; Walker & Runswick, 1993; Miroux *et al.*, 1993; Klingenberg, 1990), they were thought to span the membrane via six α-helices. Chapters 2 and 4 discuss the UCPs in more detail and attempt to increase our understanding of these proteins by predicting their 3-dimensional structure.

1.1.4 Insulin/Insulin-like signalling

As described above, there is considerable support for the role of a trade-off between the beneficial and detrimental effects of mutations in the evolution of ageing. The next im-

portant question concerns the mechanisms by which these mutations cause the damage that leads to ageing. In an attempt to address this question, various mutations have been isolated which cause an extension of lifespan. Many of these mutations, even across evolutionarily distant organisms, are found in the insulin/insulin-like signalling (IIS) pathway. This pathway is summarised in Figure 1.2.

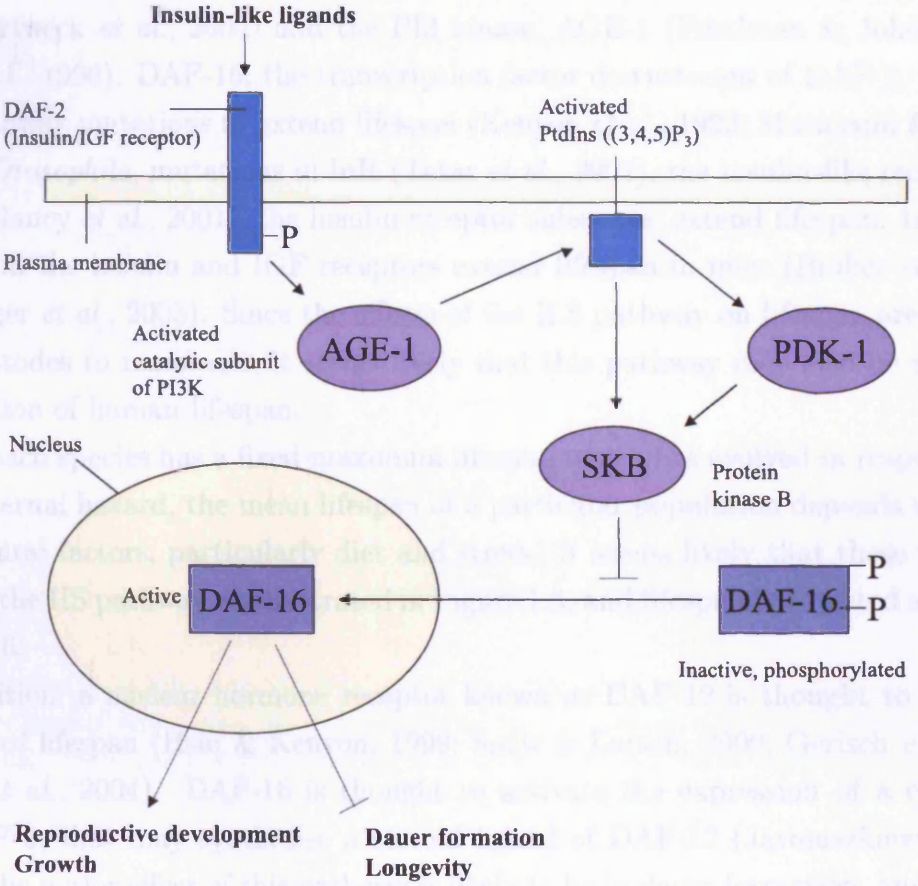


Figure 1.2: A simplified view of the insulin/IGF-like signalling pathway in *C. elegans* that regulates lifespan. Arrows indicate activation and T-bars deactivation. PI3K: phosphoinositide 3-kinase; PtdIns(3,4,5)P_3: phosphatidylinositol 3,4,5-triphosphate; PDK-1: 3-phosphatidylinositol-dependent kinase 1. This figure was drawn by hand using Microsoft PowerPoint.

In favourable conditions, insulin-like ligands bind to the insulin receptor, DAF-2, triggering its dimerisation and autophosphorylation. As a result, the catalytic subunit of phosphoinositide 3-kinase, AGE-1, dissociates from the regulatory subunit and is free to activate phosphatidylinositol. Phosphatidylinositol then in turn activates the protein kinases PDK-1 and SKB. These enzymes phosphorylate the transcription factor DAF-16, deactivating it so that it remains in the cytoplasm. In this way, genes that promote

growth, development and reproduction remain transcriptionally active. In contrast, lack of insulin-like ligands in unfavourable conditions lead to dephosphorylation and nuclear translocation of DAF-16, and genes involved in dauer formation and longevity are expressed. (The dauer is an alternative long-lived, stress resistant developmental state that *C. elegans* may adopt in unfavourable conditions, such as lack of food).

Mutations in the IIS pathway that cause lifespan extension in *C. elegans* include DAF-2, the insulin receptor (Kenyon *et al.*, 1993; Kimura *et al.*, 1997a), SKB, the Akt/PKB kinase (Hertweck *et al.*, 2004) and the PI3 kinase, AGE-1 (Friedman & Johnson, 1988; Morris *et al.*, 1996). DAF-16, the transcription factor downstream of DAF-2, is required for ILS pathway mutations to extend lifespan (Kenyon *et al.*, 1993; Murakami & Johnson, 1996). In *Drosophila*, mutations in InR (Tatar *et al.*, 2001), the insulin-like receptor, and CHICO (Clancy *et al.*, 2001), the insulin receptor substrate, extend lifespan. In addition, mutations in the insulin and IGF receptors extend lifespan in mice (Bluher *et al.*, 2003; Holzenberger *et al.*, 2003). Since the effects of the ILS pathway on lifespan are conserved from nematodes to mammals, it seems likely that this pathway may also be involved in the regulation of human lifespan.

While each species has a fixed maximum lifespan which has evolved in response to the level of external hazard, the mean lifespan of a particular population depends upon other environmental factors, particularly diet and stress. It seems likely that these factors are sensed via the IIS pathway, as illustrated in Figure 1.3, and lifespan is adjusted accordingly via DAF-16.

In addition, a nuclear hormone receptor known as DAF-12 is thought to contribute to control of lifespan (Hsin & Kenyon, 1999; Snow & Larsen, 2000; Gerisch *et al.*, 2001; Ludewig *et al.*, 2004). DAF-16 is thought to activate the expression of a cytochrome P450, DAF-9, that may synthesise a steroid ligand of DAF-12 (Jarmuszkiewicz, 2001). However, the major effect of this pathway is likely to be in dauer formation, and its effects are thought to be secondary to that of the direct action of DAF-16 in control of lifespan (Snow & Larsen, 2000; Ludewig *et al.*, 2004).

Interestingly, expression of DAF-2 in a few neurones is sufficient to prevent dauer formation (Apfeld & Kenyon, 1998). Similarly, DAF-2 or AGE-1 mutants can be rescued by restoration of IIS signalling in neurones but not muscle or intestine (Wolkow *et al.*, 2000). This suggests that DAF-2 acts non-cell-autonomously therefore a hormone is likely to be involved in the control of lifespan by IIS. Whether this hormone is the steroid ligand of DAF-12 remains to be established.

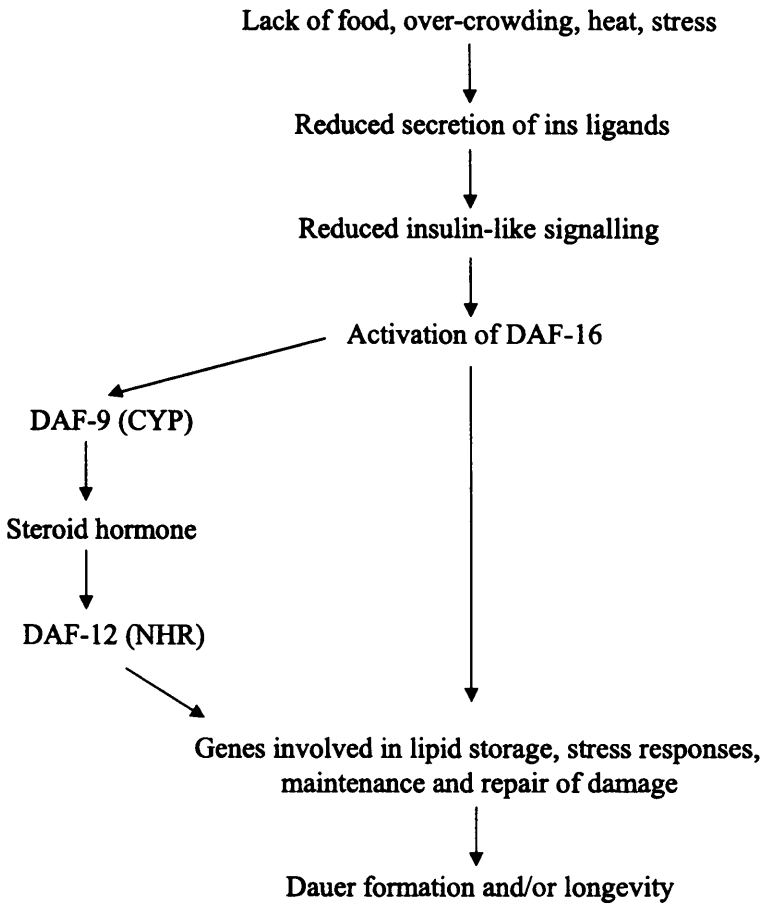


Figure 1.3: Diagram illustrating the possible mechanisms by which environmental stress may lead to a reduced rate of ageing. CYP: cytochrome P450; NHR: nuclear hormone receptor. This figure was drawn by hand using Microsoft PowerPoint.

1.1.4.1 Why does ILS reduce lifespan?

The benefit of increased lifespan but decreased fertility in response to stress may be to delay reproduction until conditions are more favourable (Masoro & Austad, 1996). For example, when food is more abundant, over-crowding is less severe or the temperature closer to optimal, reproduction is probably more likely to succeed. Dietary restriction is a method commonly used to induce unfavourable conditions experimentally, by supplying only approximately 60% of the *ab libitum* food level. In support of the idea of a reproduction/lifespan trade-off, reduced fecundity and increased lifespan are observed during dietary restriction (DR) (Zamiri, 1978; Chapman & Partridge, 1996; Chapman *et al.*, 1998; Sonntag *et al.*, 1999; Good & Tatar, 2001; Drummond-Barbosa & Spradling,

2001) and in long-lived strains of *Drosophila* (Luckinbil *et al.*, 1984; Zwaan *et al.*, 1995; Sgro & Partridge, 1999). This suggests that the effect of DR on lifespan may be linked to the cost of reproduction.

In conclusion, it is possible that a trade-off exists between fecundity and longevity. Control of ageing is likely to be mediated, at least in part, by the IIS pathway. IIS may reduce lifespan under conditions of abundant food, as a side-effect of exploiting these conditions by maximising fecundity (Partridge & Gems, 2002).

1.1.4.2 How does ILS reduce lifespan?

As described above, the genes controlled by the IIS pathway that mediate regulation of lifespan are likely to include those encoding proteins involved in repair and protection from oxidative damage and stress responses. In support of this theory, several studies have begun to identify the target proteins controlled by the ILS pathway, as these are proposed to represent the molecular determinants of ageing. These studies have found that proteins that protect against oxidative damage and other forms of stress are up-regulated in long-lived ILS pathway mutants (Honda & Honda, 1999; Holzenberger *et al.*, 2003; Hsu *et al.*, 2003; Lee *et al.*, 2003; Murphy *et al.*, 2003; Walker & Lithgow, 2003; Li *et al.*, 2004a; McElwee *et al.*, 2004). The genes down-regulated in long-lived animals are likely to be those involved with reproduction (Tissenbaum & Ruvkun, 1998) and growth (McElwee *et al.*, 2004). This area is studied in detail in Chapter 5.

1.2 Membrane proteins

Transmembrane (TM) proteins are those that span the membrane lipid bilayer. They are estimated to comprise 20-30% of all proteins (Arkin *et al.*, 1997; Wallin & von Heijne, 1998; Schwartz *et al.*, 2001; Knight *et al.*, 2004; Klein *et al.*, 2004) and are of huge biological significance since they mediate most of the communication between cells and cellular compartments. Membrane proteins fall into two classes, which span the membrane either by a bundle of α-helices (TM helices) or a barrel of β-strands (TM strands). An example of each of these classes of protein are shown in Figure 1.4. The study of TM proteins in this thesis concentrates mainly upon α-helical TM proteins.

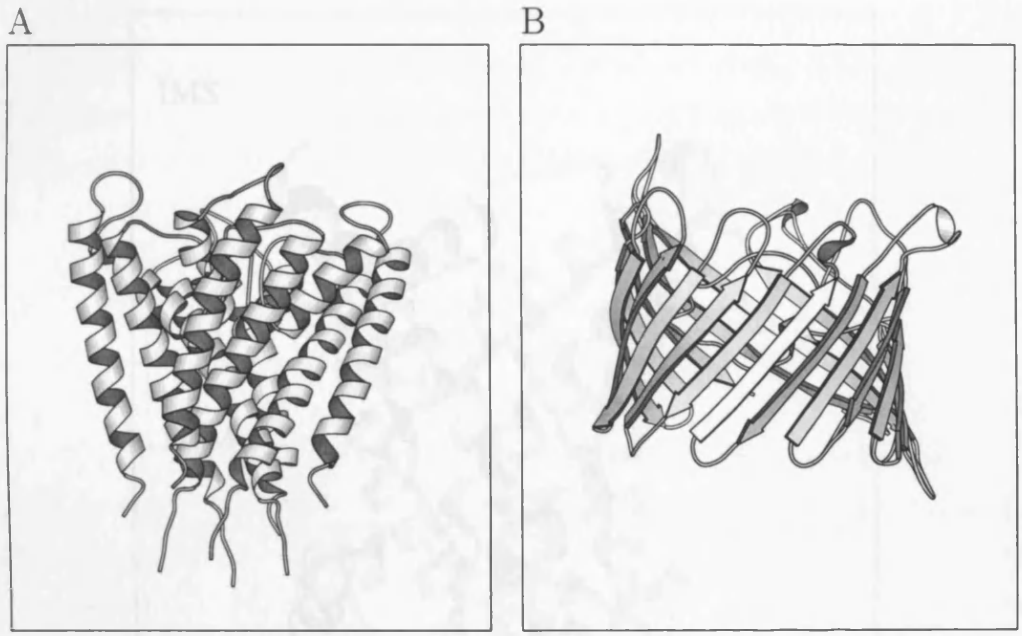


Figure 1.4: A: The structure of the Kcsa K^+ channel (1bl8) from Streptomyces lividans (Doyle *et al.*, 1998), an α-bundle TM protein. B: The structure of porin b (2por), a β-barrel TM protein (Weiss & Schulz, 1992). This figure was produced using Molscript v2.1 (© Per Kraulis, 1997-1998).

As described above, UCPs are membrane proteins with a possible role in ageing, but their mechanism of action remains very poorly understood. This lack of understanding derives mainly from the fact that, until only very recently, their 3-dimensional structure had not been determined. However, in 2003 the structure of the adenine nucleotide carrier was solved by Pebay-Peyroula *et al.* (2003) with a resolution of 2.2Å. The structure, shown in Figure 1.5, has six TM helices surrounding a pore, and shows pseudo-3-fold symmetry due to a 3-fold sequence repeat.

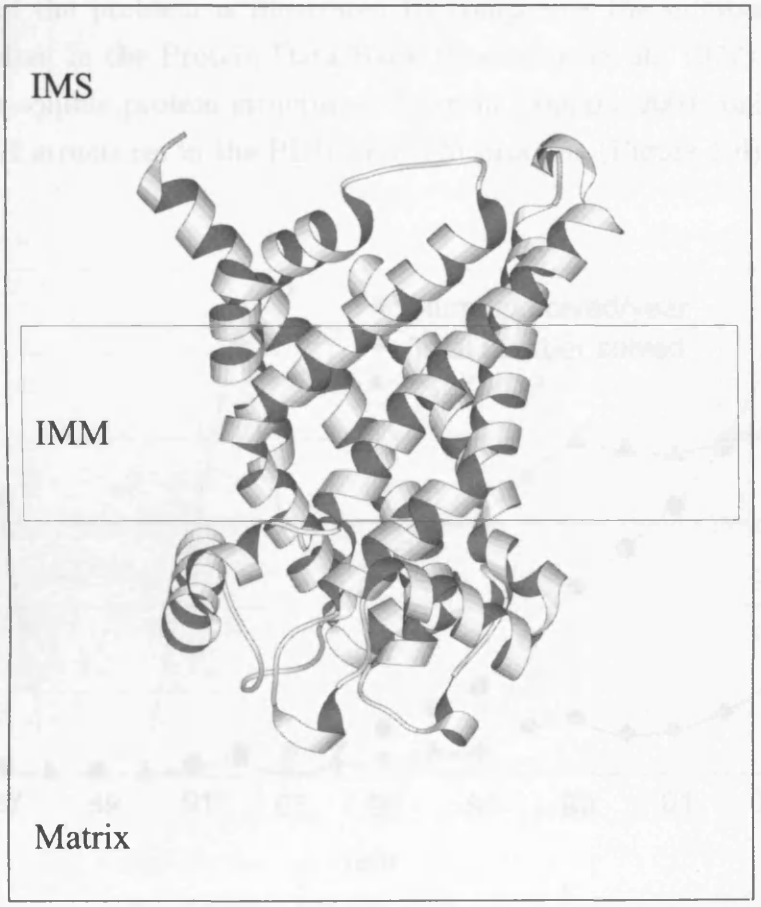


Figure 1.5: Structure of the adenine nucleotide carrier (Pebay-Peyroula *et al.*, 2003) in a view perpendicular to the membrane normal. IMS: Intermembrane space; IMM: Inner mitochondrial membrane; Matrix: Mitochondrial matrix. The red box shows the location of the membrane lipid-tail-spanning and head-group-spanning regions, as defined by PSlice (see Chapter 3). This figure was produced using MolScript.

Lack of structural information is a common problem in the study of TM proteins for two main reasons:

- Due to the small quantities of membrane proteins found in cells, and difficulties with expression systems, it is often difficult to obtain sufficient protein for analysis
- Crystallisation is often difficult to achieve because the membrane-spanning surface of membrane proteins is covered by hydrophobic residues, in order to facilitate interaction with the hydrophobic membrane. The consequent poor water solubility of TM proteins hinders their solubilisation, prior to crystallisation and structure determination.

The extent of the problem is illustrated by comparing the number of TM protein structures deposited in the Protein Data Bank (Bernstein *et al.*, 1977) (PDB) with the number of water-soluble protein structures. Even in January 2004, only approximately 0.5% of the 23792 structures in the PDB were TM proteins (Figure 1.6).

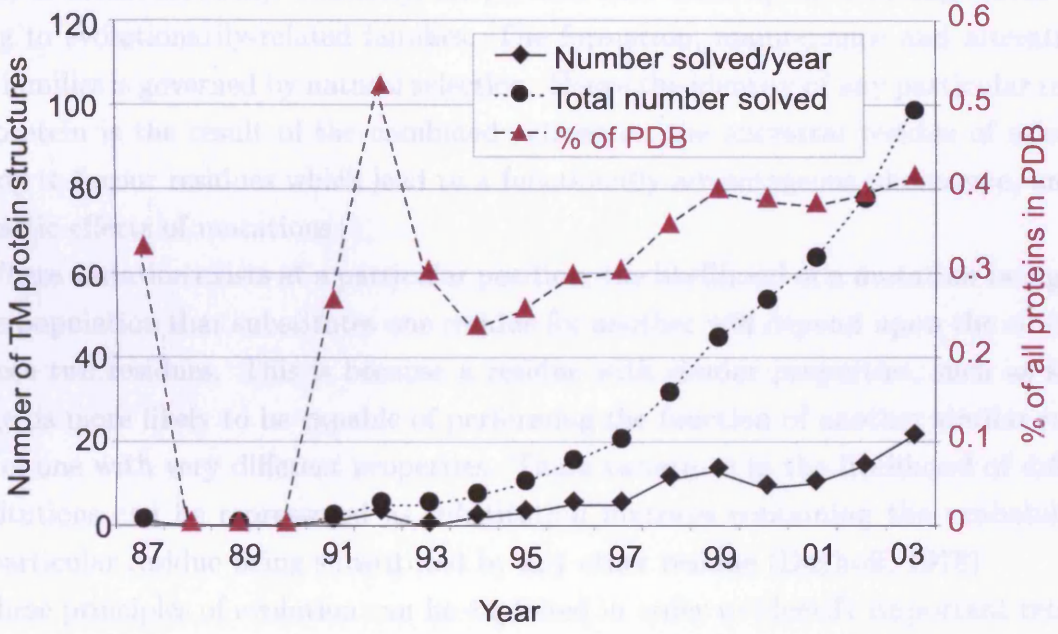


Figure 1.6: Chart illustrating the small proportion of the PDB made up of TM proteins. (All structures, including homologues, released up until the end of 2003 are included).

This lack of structural information has lead to a poor understanding of TM protein structure, compared to that for water-soluble proteins. An understanding of TM protein structure is particularly desirable since it may facilitate structural modelling until high-through-put TM protein structure determination is possible. Given the biological

importance of TM proteins such as the UCPs, this is a medically and scientifically very valuable goal. Chapters 2 to 4 of this thesis attempt to address this issue by:

- studying the available α-helical membrane protein structures in detail to increase our understanding of TM helix packing
- applying this knowledge to modelling of UCP structure by predicting the packing of its TM helices

1.3 Evolutionary theory and its application to the study of protein structure and function

Ultimately, all organisms are thought to have arisen from a universal common ancestor (Stefan & Stumvoll, 2002). Similarly, the proteins that make up the building blocks of life belong to evolutionarily-related families. The formation, maintenance and alteration of these families is governed by natural selection. Hence the identity of any particular residue in a protein is the result of the combined actions on the ancestral residue of selection, tending to favour residues which lead to a functionally advantageous phenotype, and the stochastic effects of mutations.

Where variation exists at a particular position, the likelihood of a mutation being fixed in the population that substitutes one residue for another will depend upon the similarity of those two residues. This is because a residue with similar properties, such as size or charge, is more likely to be capable of performing the function of another similar residue than of one with very different properties. These variations in the likelihood of different substitutions can be represented as substitution matrices containing the probability of any particular residue being substituted by any other residue (Dayhoff, 1978).

These principles of evolution can be exploited in order to identify important relationships between different proteins and between a protein's sequence and its function. Such techniques have been crucial throughout this thesis and are therefore described in this section.

1.3.1 Identification of homologues

Protein homologues can be identified by sequence comparison. Often this involves comparing the sequence of interest to a database of known protein sequences to identify statistically-significant relatives. A common technique, Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1990), performs this task by dividing a protein sequence

into tripeptide (three residue long) fragments. This list of fragments is then extended, by including all possible substitutions according to the mutation matrix selected, and this list is compared to a database of sequences. When a sequence from the database matches a tripeptide fragment, the overlapping regions are extended in each direction as far as possible. A score is then calculated using a mutation matrix to determine the likelihood of the substitutions suggested by that alignment. Finally, the scores of all matches are compared in order to identify the closest possible match within the database. The scores are quoted as E-values. E-values are defined as the expected number of matches with score S or greater, between a query sequence and a sequence within a database of size N.

Later, BLAST was extended to Position-Specific Iterative-BLAST, or PSI-BLAST (Altschul *et al.*, 1997). PSI-BLAST differs from BLAST in that it consists of multiple rounds of searching against the database. The first round is very similar to that performed during a BLAST search and is used to identify close relatives to the query sequence. The alignment of the query sequence and its relatives is then used to generate a position-specific score matrix, and this matrix is used to search the database in subsequent rounds. The score matrix is updated in each iteration until either no new relatives are identified or a maximum number of iterations have been performed. Profile-based methods, such as this one, are thought to be more sensitive at identifying remote sequence homologues than pairwise comparison methods (Altschul *et al.*, 1997).

Methods such as BLAST and PSI-BLAST can be of use for identifying the function of a sequence with no associated annotation. It can also identify a group of sequences belonging to a family, helping to establish phylogenetic relationships. As described below, an alignment of homologous sequences, such as that derived from BLAST or PSI-BLAST, can be used to identify conserved residues that are likely to be structurally or functionally important.

1.3.2 Detection of evolutionary sequence conservation

The calculation of residue sequence conservation amongst homologous proteins can be performed by the algorithm SCORECONS (Valdar & Thornton, 2001). SCORECONS scores each residue position of a multiple sequence alignment in terms of its conservation. The mutation matrix of Jones *et al.* (1992) is used to determine the likelihood of a particular residue being replaced by another and to calculate a score based on the variability of each position. A SCORECONS score of 0 indicates a lack of conservation at that position, whereas the maximal score of 1 indicates very high sequence conservation.

The implication of a sequence conservation score is the potential to identify residues on which natural selection has acted strongly to maintain residue identity. These residues,

with scores close to 1, are likely to have important functional or structural roles within the protein, such as ligand-binding or forming important structural contacts. This technique is employed in Chapter 4, in order to identify residues within the UCP TM helices that are likely to form contacts between the TM helices.

A similar method can also be used to identify functionally important regions of DNA sequence. This technique has been used successfully to identify transcription factor binding sites within the promoters of genes (Blanchette & Tompa, 2002; Cliften *et al.*, 2003; Berezikov *et al.*, 2004). This method, known as Phylogenetic Footprinting (Tagle *et al.*, 1988), is described in more detail in Chapter 5.

1.4 Aims of this thesis

The major goals of current ageing research are to:

- Identify the precise mechanisms by which the damage that causes ageing is produced, and how this damage leads to ageing
- Understand how the control of lifespan achieved, in response to signals from the internal and external environment
- Determine whether these mechanisms are conserved across all branches of life.

Broadly, the aim of the work within this thesis is to increase our understanding of the mechanisms of ageing. This thesis can be divided into two sections, which in turn begin to address the first two of the stated goals. Firstly, a detailed study of one specific protein with a possible role in ageing and oxidative damage is performed. This is followed by a more wide-ranging study that analyses the general mechanisms by which lifespan is controlled by ILS.

Specifically, Chapters 2-4 of this thesis are concerned with the uncoupling proteins (UCPs), α-helical transmembrane proteins with a possible role in ageing, diabetes and obesity. Chapter 2, begun in December 2001, attempts to formulate a possible model for UCP structure, using a variety of methods and published mutagenesis data.

In Chapter 3, a detailed analysis of the 24 currently available non-homologous α-helical polytopic membrane protein structures is performed, with the aim of identifying parameters with the power to predict TM helix packing. Crucially, the parameters involved are sequence-based and can therefore be used to generate structural information for proteins for which only a sequence is available. This analysis was first performed in June 2002 with 18 structures, but was updated to include a total of 24 structures in January 2004.

(While approximately 80 polytopic α-helical TM protein structures had been solved in January 2004, homologous proteins were removed to give these smaller, non-redundant datasets).

Chapter 4 aims to use the data derived from Chapter 3 to develop a method to predict transmembrane helix packing and to use this method to propose a model UCP structure. It is hoped that this work, begun in December 2002, would not only increase our understanding of UCP structures, and their possible role in ageing, but also provide a predictive method applicable to all TM protein families lacking structural information. In November 2003, after this work had been completed, the structure of a related protein, the adenine nucleotide carrier, was determined. At the end of Chapter 4, the adenine nucleotide carrier structure is compared to that of the predicted UCP model and the strengths and weaknesses of the modelling method are assessed.

Finally, Chapter 5 addresses the second goal described above: the regulation of lifespan. This work makes use of microarray data from *C. elegans* mutants with defects in the insulin-like signalling pathway, a pathway that, as described previously, has been linked to the regulation of lifespan. Data concerning the expression of genes in these mutants permits potential ageing- and longevity-promoting genes to be identified. Via the use of algorithms to identify sequence motifs that are significantly over- or under-represented in these genes, this chapter aims to identify possible transcription factor binding sites with a potential role in the regulation of lifespan.

This thesis therefore has implications for the study of membrane proteins in general, and specifically of a TM protein with a potential role in ageing. In addition, it investigates how control of lifespan is achieved through a large number of interacting regulatory proteins. It is therefore hoped that the thesis will impact upon our knowledge of ageing on a number of levels.

Chapter 2

Manual modelling of uncoupling protein structure using experimental data from the literature

2.1 Introduction

2.1.1 Aims of this chapter

The uncoupling proteins (UCPs) are a family of proteins found in the inner mitochondrial membrane that have been proposed to serve several crucial biological roles (Kerner *et al.*, 2001; Barazzoni & Nair, 2001). Whilst their primary sequence was known, and several probable functionally-important residues had been identified, very little was known of their 3-dimensional structure or mechanism of action at the time this work was carried out. The lack of structural information derived from the difficulties in practical approaches to solving the structures of transmembrane (TM) proteins due to their poor solubility in aqueous solution. The aim of this study, performed between October 2001 and May 2002, is, therefore, to gain structural information for the UCPs from a theoretical analysis of the available primary sequence and literature.

2.1.2 An overview of the uncoupling protein family

The brown fat uncoupling protein, now known as UCP1, from the golden hamster was first sequenced and its primary structure analysed by Aquila *et al.* (1985). Four paralogues are now known, UCP2, UCP3, UCP4 and brain mitochondrial carrier protein 1 (BMCP1) (Bouillaud *et al.*, 2001), the aligned sequences of which are shown in Figure 2.1. In addition, UCP3 exists in two forms, UCP3L and a truncated UCP3S, due to a partially

active polyadenylation site in intron 6 (Solanes *et al.*, 1997). The precise functions of the UCP family members remain poorly understood but various studies have suggested a role for UCPs in cold adaptation, ageing and obesity (Mizuno *et al.*, 2000; Schrauwen *et al.*, 1999; Casteilla *et al.*, 2001; Barazzoni & Nair, 2001; Kerner *et al.*, 2001).

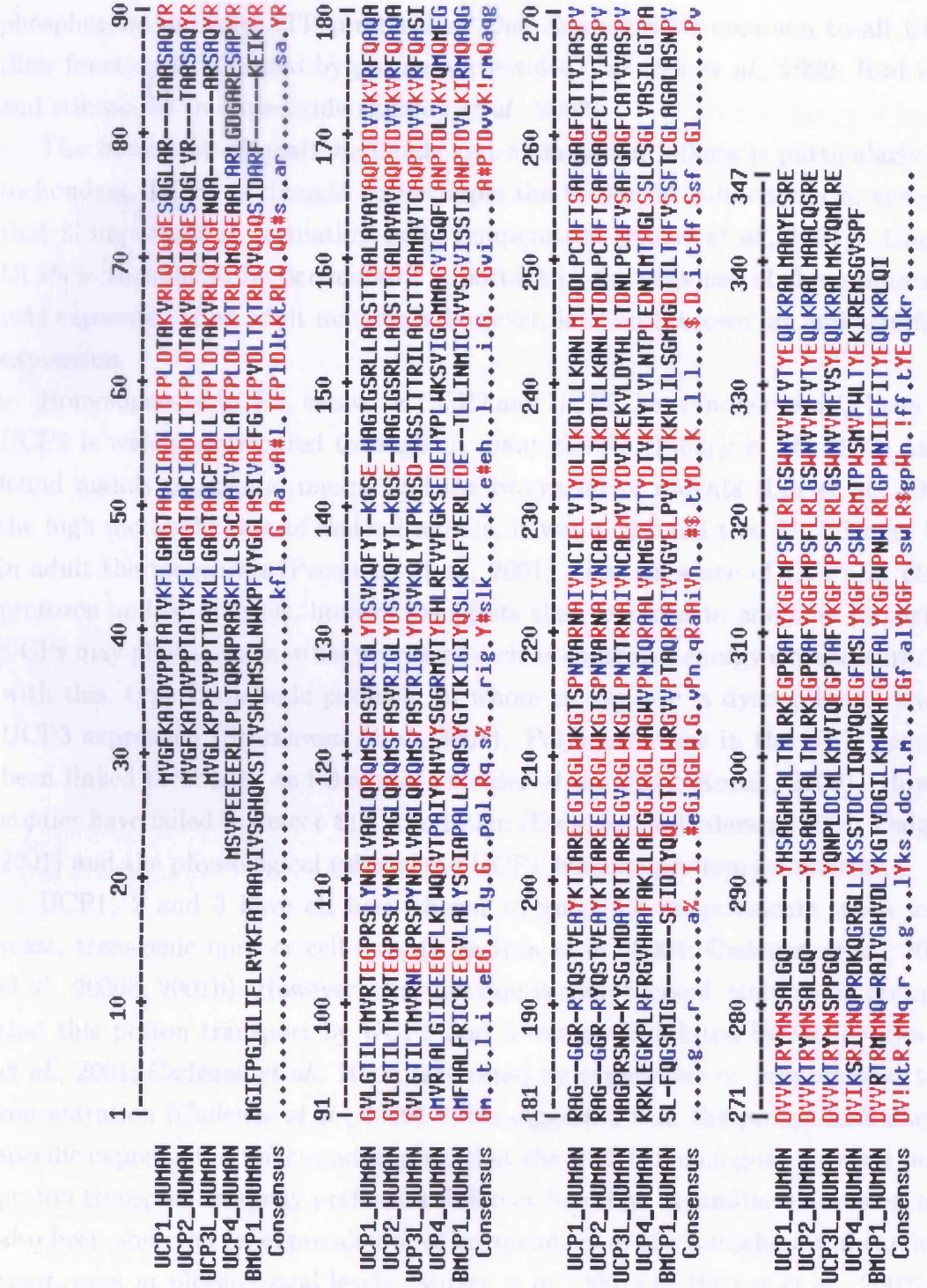


Figure 2.1: Alignment of the five human uncoupling protein paralogues. The alignment was generated using MultiAln (Corpet F, 1988). Red and blue indicate respectively positions with greater than 90% and 50% identity.

As described in Section 1.1.3.1, the UCPs are thought to catalyse leakage of protons through the inner mitochondrial membrane (IMM), back into the matrix, without ATP synthesis. Hence, as illustrated in Figure 1.1, the UCPs are said to 'uncouple' oxidative phosphorylation from ATP production. One characteristic common to all UCPs is that their function is inhibited by purine nucleotides (Jaburek *et al.*, 1999; Rial *et al.*, 1999) and stimulated by superoxide (Echtay *et al.*, 2002).

The brown fat of small mammals and mammalian infants is particularly rich in mitochondria. UCP1 is thought to uncouple the brown fat mitochondria, generating heat that is important in regulating body temperature (Klaus *et al.*, 1991). Uncoupling by UCPs is thought to be particularly important in the response of these animals to acute cold exposure. Most adult mammals, however, lack both brown fat and significant UCP1 expression.

Homologues of UCP1, known as UCP2 and UCP3, have more recently been discovered. UCP2 is widely distributed throughout many tissues (Fleury *et al.*, 1997) and UCP3 is found mainly in skeletal muscle and the brown fat of rodents (Liu *et al.*, 1998). Given the high metabolic rate of skeletal muscle, it was suggested that UCP3 may be involved in adult thermogenesis (Pecqueur *et al.*, 2001). The presence of UCPs in plants, fungi, protozoa and ectotherms, however, suggests that, perhaps in addition to thermogenesis, UCPs may play a role in other processes such as control of energy expenditure. Consistent with this, type II diabetic patients, in whom energy use is dysregulated, show reduced UCP3 expression (Schrauwen *et al.*, 2001). Polymorphisms in the UCP genes have also been linked to obesity and diabetes (Walder *et al.*, 1998; Kozak, 2000). However, some studies have failed to detect this correlation (Dalgaard & Pedersen, 2001; Dalgaard *et al.*, 2001) and the physiological roles of the UCP1 homologues remain unclear.

UCP1, 2 and 3 have all been shown to uncouple mitochondria when expressed in yeast, transgenic mice or cell lines (Kim-Han *et al.*, 2001; Cadenas *et al.*, 2002; Echtay *et al.*, 2000b, 2001b). However, this function is controversial, since some groups observed that this proton transport by UCP2 and 3 was not inhibited by nucleotides (Kim-Han *et al.*, 2001; Cadenas *et al.*, 2002), activated by superoxide or proportional to the UCP concentration (Cadenas *et al.*, 2002). This suggested that the proton leak may be a non-specific expression artefact and implied that the UCP1 homologues may not be capable of proton transport and may perform a different function. A similar expression artefact has also been observed on expression of other members of the mitochondrial carrier family in yeast, even at physiological levels (Stuart *et al.*, 2001a,b; Harper *et al.*, 2002). However, nucleotide-sensitive proton transport by all UCPs has been observed in the presence of the cofactor coenzyme Q (Echtay *et al.*, 2000b, 2001b). Hence it is likely that all UCPs are capable of proton transport that requires coenzyme Q and fatty acids and is inhibited

by nucleotides.

It has been observed that uncoupling of mitochondria by the UCPs reduces their proton motive force, and hence also the rate of generation of free radicals and other reactive oxygen and nitrogen species (ROS) (Nishikawa *et al.*, 2000; Vidal-Puig *et al.*, 2000; Arsenijevic *et al.*, 2000). This process is described in detail in Section 2.1.3.4. Uncoupling has, therefore, been proposed to reduce the rate of oxidative damage to local cellular biomolecules and hence to slow ageing (Mizuno *et al.*, 2000; Schrauwen *et al.*, 1999; Casteilla *et al.*, 2001; Barazzoni & Nair, 2001; Kerner *et al.*, 2001). It has been shown by some groups that knockout mice lacking UCPs show reduced proton leak and increased ROS production in muscle (Vidal-Puig *et al.*, 2000; Gong *et al.*, 2000), although others have been unable to confirm this (Cadenas *et al.*, 2002). Hence there is some evidence that the UCPs may be involved in the control of longevity, via the minimisation of oxidative damage.

2.1.3 Proposed physiological roles of the uncoupling protein homologues

Four principal roles have been suggested for the UCPs. These are:

- Thermogenesis
- Body weight homeostasis
- Fatty acid catabolism
- Protection from free radical damage

Evidence for and against these proposed roles will be discussed in turn. The most likely role for each UCP is summarised in Table 2.1.

2.1.3.1 Thermogenesis

The role of UCP1 in thermogenesis, via an increased proton leak across the IMM, is well established, and initially the other UCPs were also thought to be involved. Wild type mice adapt to the cold by ceasing shivering and switching to non-shivering thermogenesis, which is mediated by UCP1. UCP1 knockout mice, however, continue to shiver during prolonged periods at low temperature. This suggests that no other mechanism for thermogenesis exists, and that the other UCPs are not capable of substituting for UCP1 in this function (Golozoubova *et al.*, 2001). The physiological roles of the other uncoupling proteins therefore remain unresolved.

UCP	Role	Species transported
UCP1	Thermogenesis	Protons
UCP2	Protection from fatty acid accumulation or liberation of CoASH (Or possibly protection from free radical damage)	Fatty acids (Protons)
UCP3	Protection from free radical damage (Or possibly protection from fatty acid accumulation or liberation of CoASH)	Protons (Fatty acids)

Table 2.1: Likely roles of the UCPs and the corresponding species transported

2.1.3.2 Body weight homeostasis

Clapham *et al.* (2000) found that mice over-expressing UCP3 are small, despite hyperphagia (increased food consumption). This suggested a role for UCP3 in the control of energy usage. Consistent with this role, UCP2 and UCP3 mRNA expression is up-regulated 2-fold during a high fat diet (Bezaire *et al.*, 2001). However, UCP3 knockout mice are not obese (Gong *et al.*, 2000), indicating that this role is not clear-cut or direct.

2.1.3.3 Fatty acid catabolism

In addition to up-regulation during a high fat diet, UCP2 and UCP3 mRNA expression is increased 4-fold during fasting (Bezaire *et al.*, 2001). The common aspect of each of the conditions that stimulate UCP expression (cold exposure, fasting and a high fat diet) is an increased fat catabolism. This suggests that the primary role of the UCP2 and 3 may be linked with this process. Consistent with this role, UCP3(-/-) mice show respiratory changes that are associated with reduced fatty acid catabolism (Bezaire *et al.*, 2001).

This work suggests that the action of UCP2 and 3 may be not proton translocation but export of fatty acids from the mitochondrial matrix. Two particular roles for this export function have been proposed: (1) to facilitate rapid β-oxidation by allowing CoASH to return to the usable pool and (2) to protect against the accumulation of fatty acids in the matrix.

During fat catabolism, fatty acids enter the mitochondrion as acyl-CoA. Mitochondrial thioesterase 1 (MTE-1) cleaves acyl-CoA, liberating CoASH and fatty acid anions within the mitochondrial matrix. The physiological role of this process is unknown, but

some groups believe that UCP3 then transports the fatty acid anions back out of the mitochondrion in order to their prevent accumulation in the matrix, where they cannot be metabolised (Himms-Hagen & Harper, 2001). This would explain the up-regulation of UCP3 during conditions in which fat oxidation is high, since at these times the acyl-CoA levels would be highest and the cell would be at greatest risk from fatty acid anion accumulation. In addition, the increased demand for CoASH would favour a mechanism in which this molecule was liberated from existing complexes for re-use. Support for this role for UCP3 and MTE-1 comes from the observed co-ordinated regulation of their expression during conditions in which the plasma lipid concentration is altered (Clapham *et al.*, 2001). In contrast, there appears to be no correlation between the expression patterns of MTE-1 and UCP1 or UCP2, suggesting that these have a different physiological role.

2.1.3.4 Protection from free radical damage

The primary role of the uncoupling protein homologues may be to minimise production of reactive oxygen species, in order to protect the cell from free radical damage. ROS production is thought to occur by the following mechanism: When the demand for ATP is low, little adenosine diphosphate will be present for phosphorylation by ATP synthase. As a result, fewer protons will be used by ATP synthase and so the proton motive force will increase. The greater proton gradient will lead to more energy being required for the electron transport chain to transfer protons across the IMM. The flow of electrons will therefore be slower, so that each of the electron transport chain intermediates will remain reduced (carrying an extra electron) for longer. Certain intermediates, notably semi-quinone, will be in radical form when reduced, and their increased lifespan leads to an increased likelihood of reaction with oxygen. On reaction of an electron with oxygen a superoxide radical, $O_2^{\cdot -}$, is formed.

The superoxide radical, although relatively unreactive and unable to cross the mitochondrial membrane itself, is capable of reaction with a number of other species, so forming more reactive radicals which are thought to cause significant damage to various cellular macromolecules. Radicals, being in general highly reactive, are unable to travel long distances without reaction. Hence the primary targets are thought to be molecules in close proximity to the site of radical generation, such as the mitochondrial DNA. The action of the UCPs will tend to reduce the proton motive force, and hence also the tendency of the electron transport chain intermediates to react with oxygen and to produce radicals. Since uncoupling by the UCPs is stimulated by superoxide (Echtay *et al.*, 2002), the UCPs have been proposed to form an essential part of the electron transport chain, modulating the cellular levels of superoxide, and consequently of ROS (Casteilla *et al.*,

2001). The widespread distribution of the UCPs, in plants, animals, fungi and protozoa, suggests that they represent a general strategy for maintenance of redox balance.

Much evidence has been obtained in support of the view that UCPs may protect against ROS damage. For example, mitochondrial superoxide production can be prevented by relatively modest levels of UCP1 expression (Nishikawa *et al.*, 2000), or increased by knockout of UCP3 in mouse skeletal muscle (Vidal-Puig *et al.*, 2000). Similarly, the macrophages of UCP2 knockout mice generate greater quantities of ROS (Arsenijevic *et al.*, 2000). It has been proposed by Kim-Han *et al.* (2001) that in the brain this action may allow the brain uncoupling protein, BMCP1, to protect neurones against the oxidative damage that is thought to be a major cause of the pathophysiological changes that occur in neurodegenerative conditions. The group found that, while uncoupling was increased, superoxide production was 25% lower in cells weakly over expressing BMCP1 compared to controls. Using mitochondrial depolarising agents they showed that this corresponds to the majority of mitochondrially derived superoxide.

Work using transgenic knockouts and mice over-expressing UCP3 has indeed shown that, while UCP3 does not contribute to basal proton conductance or basal metabolic rate, it does seem to catalyse a superoxide-induced proton conduction (Echtay *et al.*, 2002). Although they observed a 4-fold increase in proton conductance in transgenic mice over-expressing UCP3, Cadenas *et al.* (2002) proposed that this was an artefact not likely to be due to any native uncoupling activity of UCP3, for two reasons: (1) The increase in proton conductance observed was not proportional to the increase in UCP3 expression. (2) The increased conductance did not share the properties of induction by superoxide and inhibition by purine nucleotides that is characteristic of native uncoupling by the UCPs (Cadenas *et al.*, 2002). However, in UCP3 knockout mice, superoxide-induced proton conductance was abolished without affecting the basal metabolic rate, suggesting this inducible conductance is the physiological role of UCP3 (Echtay *et al.*, 2002). This suggests that the UCPs may form part of a ROS defence system that decreases ROS production when radicals accumulate, operating a compromise between loss of energy efficiency during uncoupling and increased risk of ROS damage in the absence of uncoupling.

2.1.4 Proposed role of the UCPs in ageing

UCPs have been proposed to form an essential part of the electron transport chain, modulating the production of superoxide, and consequently of ROS (Casteilla *et al.*, 2001). An increased ROS burden is believed to cause cellular damage that contributes to ageing. This is known as the Oxidative Damage Theory of ageing (see Section 1.1.3). Evidence,

from studies over expressing antioxidant enzymes, is accumulating in favour of this hypothesis (Parkes *et al.*, 1998b; Hari *et al.*, 1998; Schwarze *et al.*, 1998; Parkes *et al.*, 1998a; Sun & Tower, 1999).

Some interesting studies have suggested that UCPs may play a role in the increased ROS damage that occurs during ageing. Firstly, both reduced expression levels of UCP3 and fewer mitochondria are found within skeletal muscle of aged rats (Kerner *et al.*, 2001; Barazzoni & Nair, 2001), although one study has failed to find any change in numbers of mitochondria (Farrar *et al.*, 1981). This reduced UCP3 expression is accompanied by the expected decrease in uncoupling. There are no changes in normal respiration or in the activity of other mitochondrial enzymes, indicating that mitochondrial efficiency in general is not impaired (Kerner *et al.*, 2001). The decrease in number of mitochondria is likely to contribute to the weakness that occurs concurrently with loss of muscle mass in the elderly. However, perhaps it is necessary, in order to partially compensate for the increased production of ROS caused by reduced UCP3 expression. Alternatively, perhaps UCP expression is lowered as a result of decreased demand, when less ROS are produced from the lower numbers of mitochondria present. More work is needed to establish these cause/effect relationships in detail.

Interestingly, it has been proposed that UCPs may mediate the protective effects of some polyunsaturated fatty acids against the age-associated diseases (Cha *et al.*, 2001). These lipids stimulate UCP2 and UCP3 expression via the peroxisome proliferator-activated receptors (PPARs) when fed to mice (Armstrong & Towle, 2001; Cha *et al.*, 2001). This mechanism may be responsible for the action of polyunsaturated fats to reduce obesity and as a result protect against cardiovascular disease. Polyunsaturated fatty acids also have many other physiological effects which may be involved in, or entirely responsible for, these actions and further work is needed to either confirm or rule out a role for the UCPs.

2.1.5 Location and regulation of UCP expression

Information concerning the distribution and control of UCP expression is given in Tables 2.2 and 2.3.

2.1.6 The structural organisation of the uncoupling proteins

The UCPs each have a molecular weight of approximately 33kDa and a length of 306-309 amino acids. According to our current understanding, their main features are:

- Three highly homologous tandem domains

UCP	Location	Reference
UCP1	Brown fat and longitudinal smooth muscle	Nibbelink *et al.* (2001)
UCP2	Many tissues	Fleury *et al.* (1997)
UCP3	Skeletal muscle and brown fat of small mammals	Liu *et al.* (1998)
UCP4	Brain	Mao *et al.* (1999)
BMCP1	Many tissues particularly CNS	Sanchis *et al.* (1998)

Table 2.2: Tissue expression patterns of the UCPs

- Six transmembrane α-helices with both N and C termini located towards the cytosol
- A nucleotide-binding domain, to which nucleotides bind and regulate UCP activity

These features are discussed in detail below.

2.1.6.1 The tripartite structure of the uncoupling proteins

The UCPs consist of 3 homologous domains, each approximately 100 amino acids long (Aquila *et al.*, 1985). All three domains show significant similarity to the mitochondrial carrier protein signatures, as defined by various protein family databases (PFAM family PF00153; PRINTS PR00784; PROSITE PS00215 and INTERPRO IPR001993). The UCPs are, therefore, members of the mitochondrial carrier protein family (MCF), along with the ADP/ATP translocase, the 2-oxoglutarate/malate carrier, the mitochondrial phosphate carrier and many others. The members of this family of proteins are evolutionarily membrane proteins found in the inner mitochondrial membrane, across which they transport a variety of substances.

The UCPs themselves form the INTERPRO IPR002030 subfamily, known as the mitochondrial brown fat uncoupling proteins, within the MCF. Hence their sequence contains regions conserved among the UCPs alone, and among the whole mitochondrial carrier family, as shown in Figure 2.2.

Homology between the three domains suggests that the present UCPs were formed by triplication of an ancestral 100 residue domain. Analysis of conserved residues highlights this tripartite structure, since the same residues are often found at equivalent positions in each domain. Many of these residues are characteristic of the mitochondrial carrier

Factor	Reference	Notes
Oestrogen	Pedersen *et al.* (2001)	UCP1 is up-regulated in brown adipose tissue and UCP2 in white adipose tissue. May contribute to the effects of oestrogen on body weight and energy expenditure associated with ovarectomy.
IGF-1	Gustafsson *et al.* (2001)	IGF-1 implicated in protection of neurones from ROS damage. Loss of IGF-1 in Non Insulin Dependent Diabetes may lead to associated neuropathies via decreased UCP3.
PPAR α and γ	Kelly *et al.* (1998)	UCP1, 2 and 3 are all up-regulated.
Beta-3 agonists	Gong *et al.* (1997)	May mediate response to cold. Observed for UCP1 and UCP3 but not UCP2.
Retinoids	Alvarez *et al.* (1995)	-
Thyroid hormone	Larkin *et al.* (1997); Gong *et al.* (1997); Branco *et al.* (1999)	Observed for UCP1 and UCP3 but not UCP2.
Leptin	Gong *et al.* (1997)	-

Table 2.3: Factors stimulating the expression of the UCPs. In addition, all UCPs are inhibited by purine nucleotides, and both fatty acids and coenzyme Q are required for their activity.

family. Three arginine residues, for example, are found at the same point in TM helix 2, 4 and 6 (residues R83, R182 and R276 in UCP1).

2.1.6.2 Predicted transmembrane organisation of the uncoupling proteins

Confirming the work of Aquila *et al.* (1985), hydropathy analysis of human UCP1 and adenine nucleotide carrier suggests that they have 6 transmembrane helices, although TMs 2 and 4 are likely to be amphipathic and are more weakly detected. The six regions of increased hydrophobicity can be seen on a hydropathy plot, as shown in Figure 2.3. The reduced hydrophobicity of TMs 2 and 4 relative to the others may have important implications for the modelling of the UCPs, and will be investigated further in later work. Specifically it may suggest that these helices are located within the helix bundle, rather than adjacent to membrane lipid-tails. As illustrated in Figure 2.4, the six transmembrane

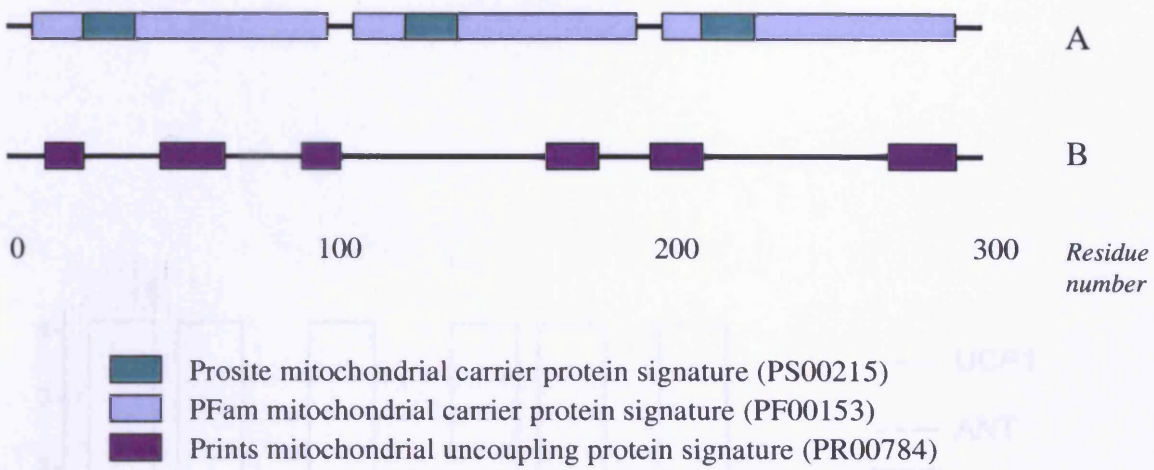


Figure 2.2: Schematic diagram showing the domain organisation of the uncoupling proteins. A: PROSITE and PFAM signatures found in all mitochondrial carrier proteins. B: PRINTS Signature unique to the UCP subfamily.

helices are each encoded by a separate exon of the UCP genes (Kozak *et al.*, 1988).

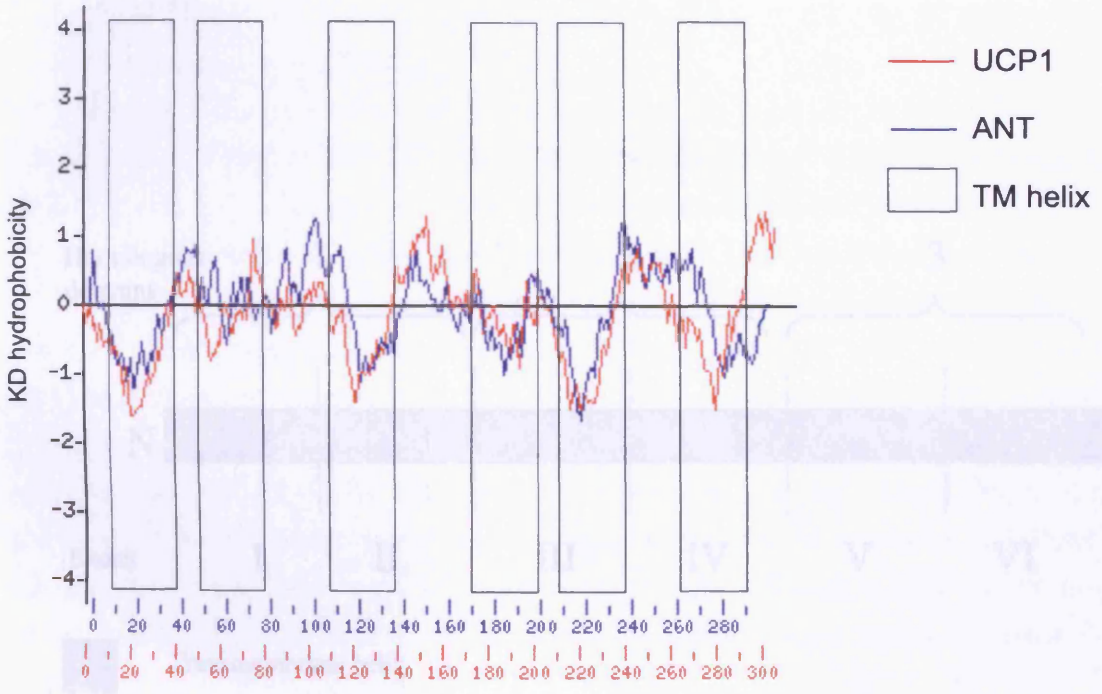


Figure 2.3: Hydropathy analysis of human UCP1 and adenine nucleotide carrier showing predicted transmembrane regions. The plot was produced using the Weizmann Bioinformatics Server (*http://bioinformatics.weizmann.ac.il/hydroph/*), using the Kyte and Doottlittle hydrophobicity scale (Kyte & Doolittle, 1982) and a window size of 19. Hydrophobicity values below 0 indicate high hydrophobicity and a likely TM region.

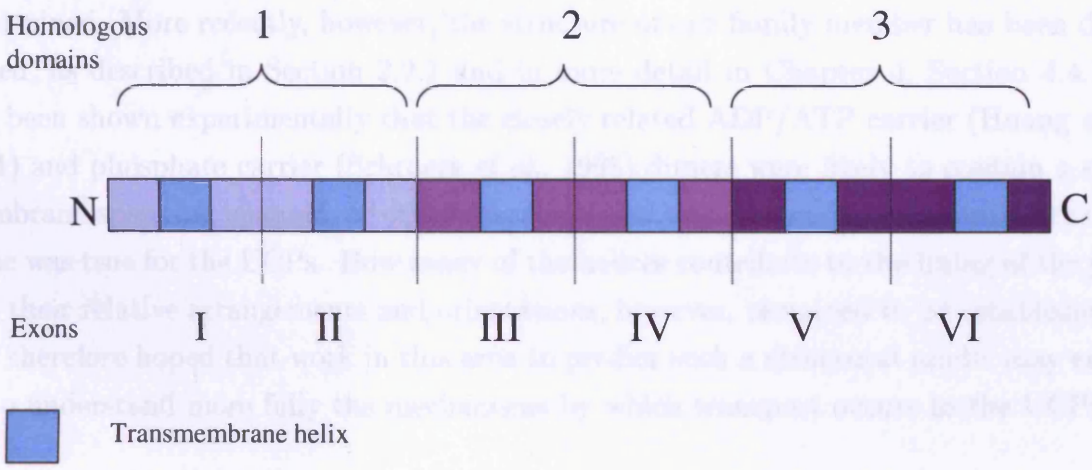


Figure 2.4: Schematic diagram showing the exons and tripartite structure of the uncoupling proteins, and the locations of the transmembrane helices.

Various groups have used antibodies, tryptic cleavage and affinity labelling reagents to probe the organisation and orientation of the UCPs in the membrane. Regions shown to be on the matrix side include residues 61-79, 164-184 and 253-279 (Miroux *et al.*, 1993, 1992) and on the cytosolic side the N and C termini and residues 105-118 (Eckerskorn & Klingenberg, 1987; Miroux *et al.*, 1993). In addition, Gonzalez-Barroso *et al.* (1997) have identified matricial loops in each domain that come together to control transport through the protein by acting as a gate. Removal of this region allows the UCP to function as a non-specific pore permeable to molecules up to 1kDa, rather than a proton-specific carrier. These gate regions are also thought to be involved in the binding of purine nucleotides (Gonzalez-Barroso *et al.*, 1999).

The view of the structure of the UCP family at the time the present study was carried out is shown in Figure 2.5. The three dimensional arrangement of the transmembrane helices of the UCPs, or any member of the mitochondrial carrier family, had yet to be determined. More recently, however, the structure of one family member has been determined, as described in Section 2.2.1 and in more detail in Chapter 4, Section 4.4.1. It had been shown experimentally that the closely related ADP/ATP carrier (Huang *et al.*, 2001) and phosphate carrier (Schroers *et al.*, 1998) dimers were likely to contain a single membrane-spanning channel, or other transport pathway. Hence it was assumed that the same was true for the UCPs. How many of the helices contribute to the lining of the pore, and their relative arrangements and orientations, however, remained to be established. It was therefore hoped that work in this area to predict such a structural model may enable us to understand more fully the mechanisms by which transport occurs in the UCPs.

2.1.6.3 The proposed purine nucleotide-binding domain

Like the related ADP/ATP carrier, another member of the MCF, PROSITE similarity analysis suggests that the C terminal regions of the UCPs contain a purine nucleotide-binding domain. Whereas the nucleotide is the substrate for transport by the ADP/ATP carrier, the UCPs are inhibited by purine nucleotides (Jaburek *et al.*, 1999; Rial *et al.*, 1999). This site is important in the control of UCP function and hence, perhaps, also in processes such as energy metabolism and ageing.

The presence of such highly conserved charged residues within the transmembrane helices (shown in Figure 2.5) initially highlighted that they may play important functional roles. It now seems that many are involved in nucleotide binding. Whilst the binding site is located mainly in the third domain of the protein, packing of the helices allows contribution from homologous residues in other domains. For example, the three conserved arginine residues, R83, R182 and R276, are thought to bind the phosphate groups

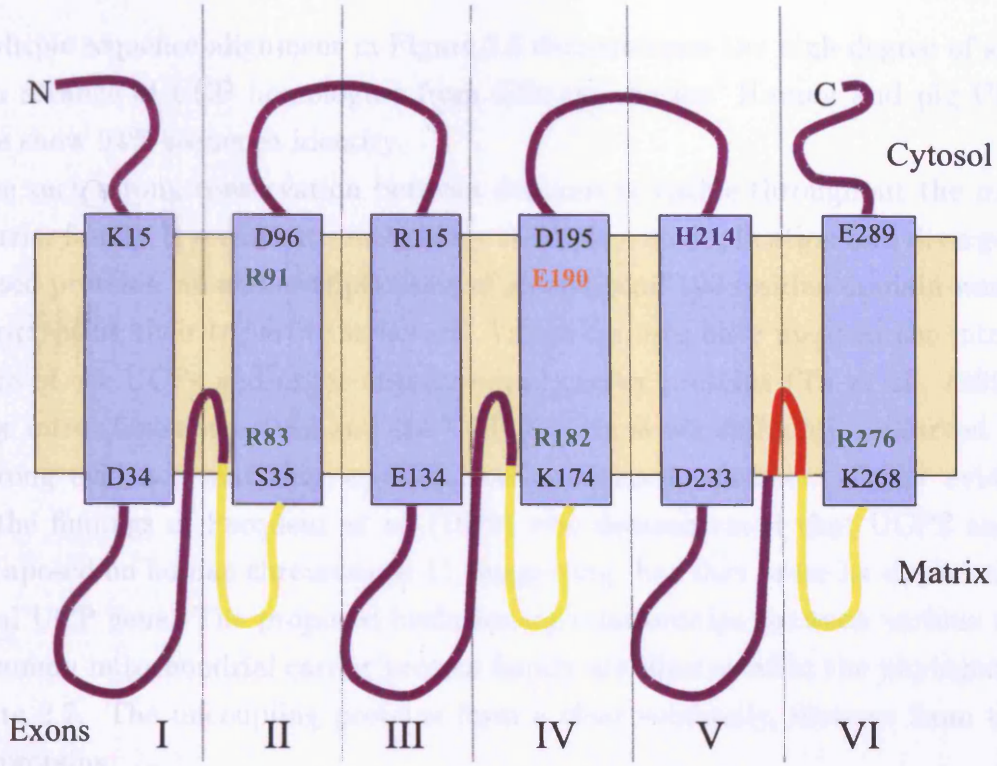


Figure 2.5: Proposed structure of UCP1 as of October 2001, highlighting important functional residues: Black- Helix termini; Red- Purine nucleotide ring binding; Green- Nucleotide phosphate group binding; Blue- Regulation of NTP (but not NDP) binding; Yellow- Gating loops forming hydrophobic pocket for binding of purine moiety of nucleotide; Orange- pH sensor for nucleotide-binding. Each exon is indicated by a Roman Numeral and separated from the others by a vertical black line.

of the inhibitory purine nucleotide. This was shown when replacing them with uncharged residues abolished nucleotide-binding in isolated UCP1 (Echtay *et al.*, 2001a).

Nucleotide binding is pH-dependent. This property is thought to be provided by protonation of the carboxyl group of E190 (Winkler *et al.*, 1997). In support of this idea, removal of this carbonyl group by mutation of glutamic acid 190 to glutamine gives the predicted mutant in which binding affinity is no longer affected by pH (Echtay *et al.*, 1997). Site directed mutagenesis has shown that residues H214, D209 and D210 also contribute to the pH-dependence of binding. H214 binds to NTP but not NDP in a pH-dependent manner (Echtay *et al.*, 1998) and the two adjacent aspartates are thought to function by holding H214 in the correct orientation in the binding site (Echtay *et al.*, 2000a).

2.1.6.4 Evolutionary conservation

The multiple sequence alignment in Figure 2.6 demonstrates the high degree of similarity between a range of UCP homologues from different species. Human and pig UCP2, for example show 94% sequence identity.

Since such strong conservation between domains is visible throughout the mitochondrial carrier family, it seems extremely likely that, prior to duplication and divergence into specialised proteins, an earlier triplication of an ancestral 100 residue domain was responsible for creating their tripartite structure. Various groups have mapped the intron/exon structure of the UCPs and other mitochondrial carrier proteins (Tu *et al.*, 1999). That both the intron/exon structure and the UCP sequences are so highly conserved provides very strong evidence that they evolved from a common ancestor. Other evidence includes the findings of Pecqueur *et al.* (1999) who demonstrated that UCP2 and UCP3 are juxtaposed on human chromosome 11, suggesting that they arose by duplication of an ancestral UCP gene. The proposed evolutionary relationships between various members of the human mitochondrial carrier protein family are illustrated in the phylogenetic tree in Figure 2.7. The uncoupling proteins form a clear subfamily, distinct from the other carrier proteins.

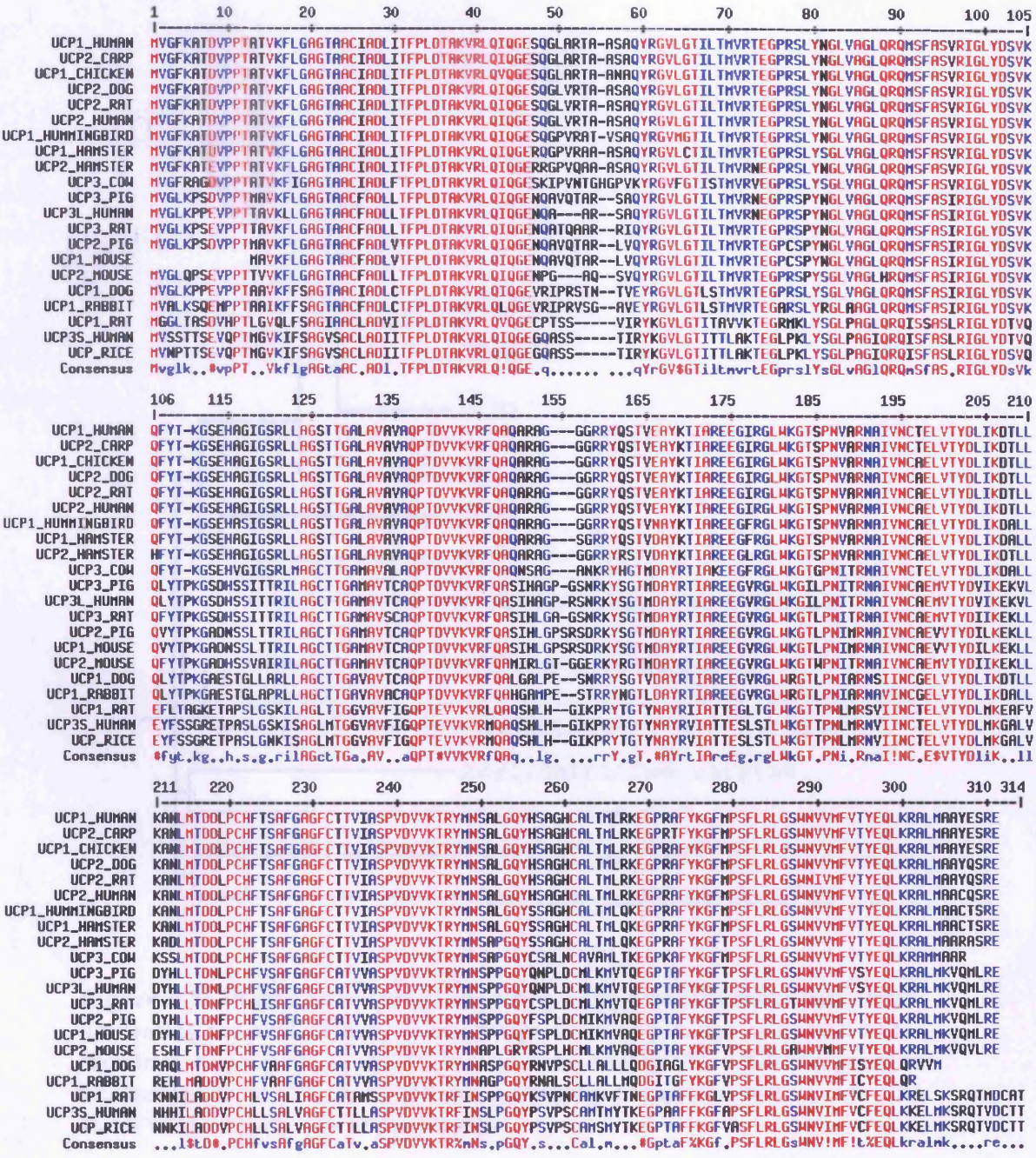


Figure 2.6: Multiple sequence alignment of various UCPs demonstrating their high sequence identity. The alignment was generated using MultiAln (Corpet F, 1988). Red and blue indicate respectively positions with greater than 90% and 50% identity.

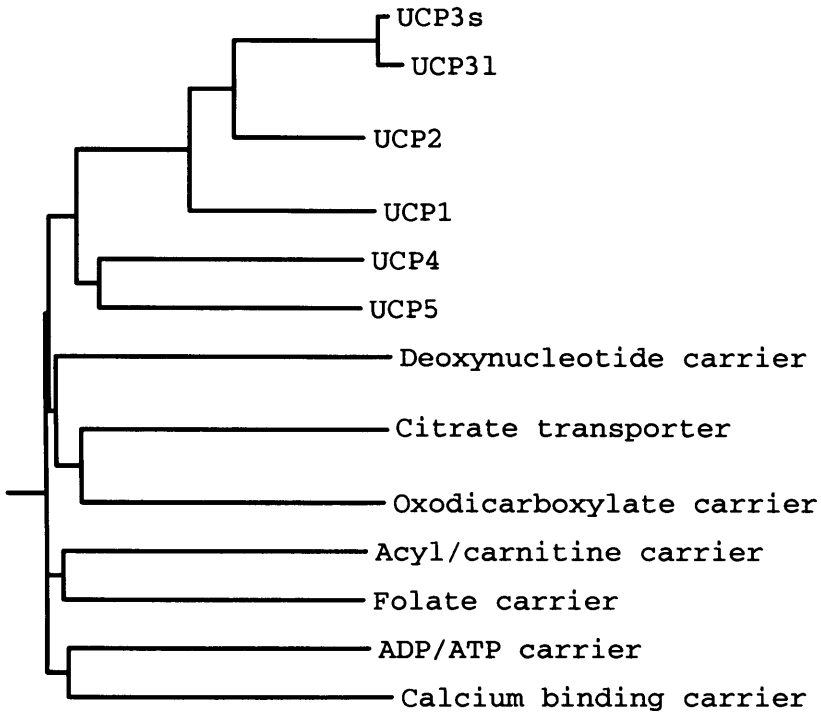


Figure 2.7: Phylogenetic tree produced by Phylip v3.5c (Felsenstein, 1993), illustrating the evolutionary relationships between a range of human mitochondrial carrier proteins. UCP5 is also known as BMCP1. UCP3s represents a truncated form of UCP3l. All default Phylip parameters were used. The scales used on each axis are arbitrary.

Figure 2.8 illustrates the major sequence characteristics of the UCPs and other members of the mitochondrial carrier family, as defined by visual inspection of the aligned sequences. It can generally be assumed that features conserved across a range of species are likely to have a functional or structural role which places evolutionary constraints on them. The UCPs can be distinguished from each other, and from the other members of the mitochondrial carrier family, by various conserved signatures. Some conserved residues are present throughout the whole carrier protein family, for example, the conserved arginines, R83, R182 and R276, whilst others, such as the conserved aspartate in the first transmembrane helix, D27, are found only in the uncoupling proteins. These conserved residues are likely to contribute to the function of the protein and the differences in these residues between homologues are likely to indicate differences in function between them.

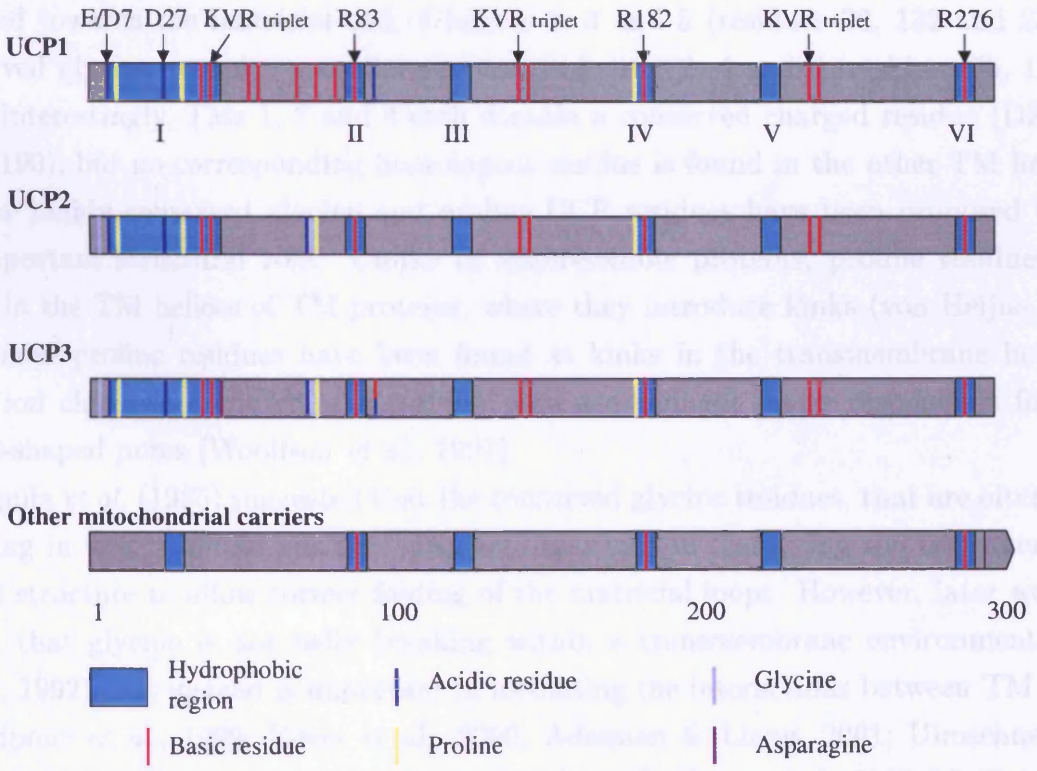


Figure 2.8: Cartoon showing the conserved features of the sequences of mitochondrial carrier protein family members as defined by visual inspection of a sequence alignment. Highly conserved residues are indicated with arrows and transmembrane helices by Roman numerals.

In general, UCP1, 2 and 3 can easily be distinguished from the sequences of the other, much less highly conserved, mitochondrial carrier proteins. UCP4 and BMCP1 (UCP5),

however, are much more similar in sequence to the other carriers and lack the signatures of the other UCPs. Consistent with this, in the human, UCP2 and UCP3 both show 59% identity with UCP1 (Fleury *et al.*, 1997; Liu *et al.*, 1998), whereas UCP4 and BMCP1 are more distantly related, with 34% and 30% identity respectively to UCP1 (Sanchis *et al.*, 1998; Mao *et al.*, 1999). There is in fact no evidence that UCP4 and 5 share uncoupling activity with UCPs 1-3. The discussion within this chapter applies primarily to UCP1, 2 and 3, since these can be taken as the archetypal uncoupling proteins.

The UCP sequence can be divided into 6 exons, each approximately 50 residues in length and each containing a relatively hydrophobic stretch that forms a transmembrane helix. No single domain is considerably more evolutionarily conserved than the others.

As would be expected, the tripartite structure of the UCPs is strongly reflected in their pattern of conserved residues. For example, conserved arginines are found at equivalent positions in the second, fourth and sixth transmembrane helices (R83, R182, R276) of all members of the mitochondrial carrier protein family. Similarly, a conserved proline is found towards the matricial end of helices 1, 3 and 5 (residues 32, 132 and 231). A conserved glycine occupies a similar position in helices 2, 4 and 6 (residues 76, 175 and 269). Interestingly, TMs 1, 2 and 4 each contain a conserved charged residue (D27, R91 and E190), but no corresponding homologous residue is found in the other TM helices.

The highly conserved glycine and proline UCP residues have been proposed to play an important structural role. Unlike in water-soluble proteins, proline residues often occur in the TM helices of TM proteins, where they introduce kinks (von Heijne, 1991). Conserved proline residues have been found at kinks in the transmembrane helices of other ion channels (Schiffer *et al.*, 1995), and are thought to be responsible for their funnel-shaped pores (Woolfson *et al.*, 1991).

Aquila *et al.* (1985) suggested that the conserved glycine residues, that are often helix-breaking in water-soluble proteins, may be important in disrupting the transmembrane helical structure to allow correct folding of the matricial loops. However, later work has shown that glycine is not helix breaking within a transmembrane environment (Li & Deber, 1992), but instead is important in mediating the interactions between TM helices (Javadpour *et al.*, 1999; Eilers *et al.*, 2000; Adamian & Liang, 2001; Ulmschneider & Sansom, 2001), often within GxxxG and related motifs (Senes *et al.*, 2000; McClain *et al.*, 2003; Whittington *et al.*, 2001; Mendrola *et al.*, 2002; Liu *et al.*, 2002; Arselin *et al.*, 2003; Sulistijo *et al.*, 2003; Overton *et al.*, 2003; Polgar *et al.*, 2004; Kairys *et al.*, 2004; Li *et al.*, 2004b; Lee *et al.*, 2004). Several glycine residues in the UCPs appear to be well conserved throughout the uncoupling proteins but poorly in the other members of the mitochondrial carrier family. This provides further support for the idea that the conserved glycines are not required for transmembrane helix termination, since all members of the mitochondrial

carrier protein family share a similar transmembrane organisation. Hence these conserved glycine residues are likely to have a role in mediating the interactions of TM helices that is in some way specific to the UCPs.

Aromatic residues are another group that may be used to make inferences about TM protein structure. These residues tend to be conserved between family members near the interface of the TM lipid-tails and head-groups, probably with a role in anchoring (Killian & von Heijne, 2000; Yuen *et al.*, 2000; Ulmschneider & Sansom, 2001), and can therefore in theory be used to infer the approximate location of this interface. As can be seen in Figure 2.9, the UCPs contain a number of aromatic residues in positions that are highly consistent with the predicted locations of their TM helices.

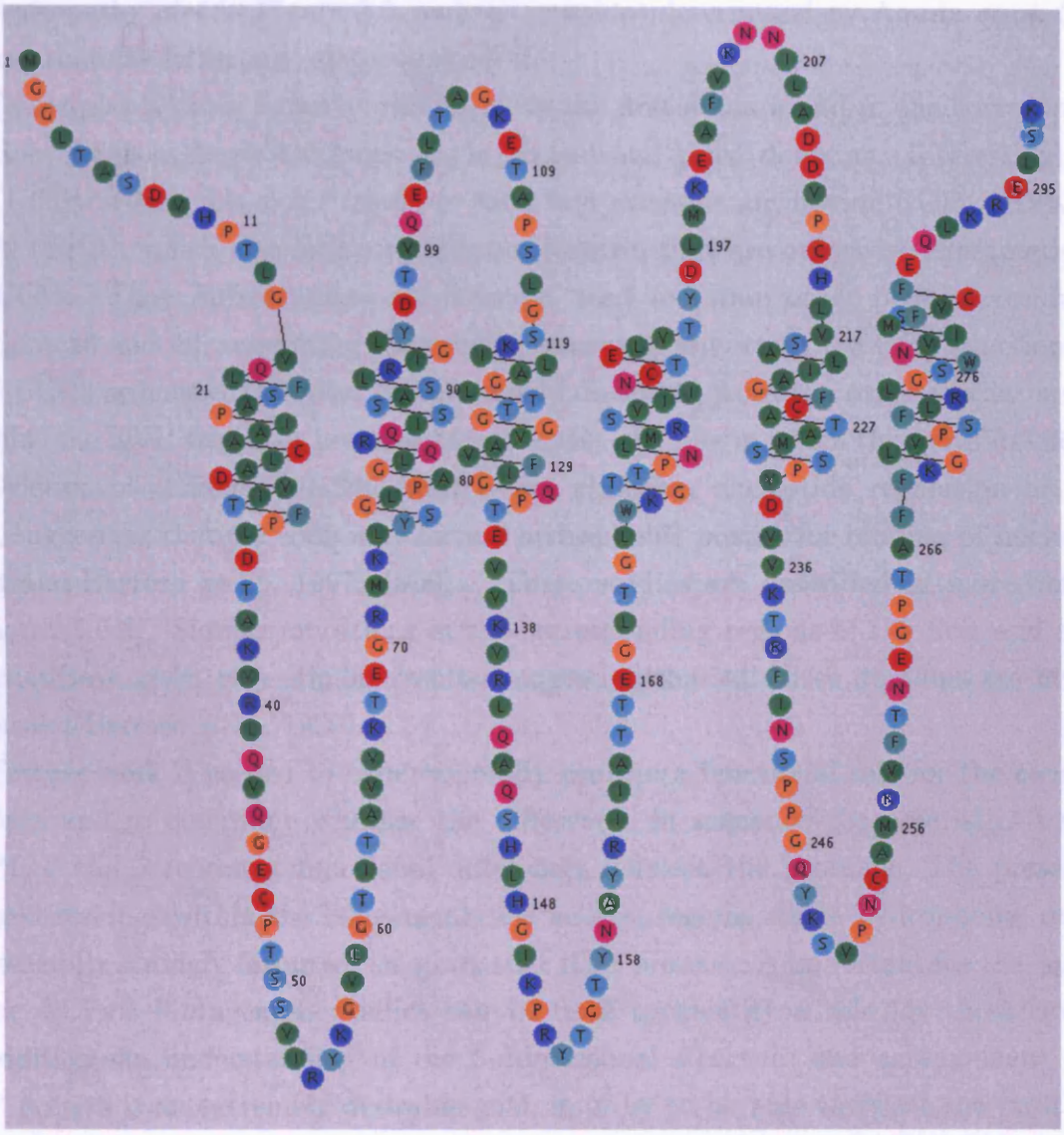


Figure 2.9: Residue-based diagram of UCP1. The locations of TM helices are taken from Aquila *et al.* (1985). This figure was produced using the Residue-based Diagram Editor (RbDe) web service (Campagne & Weinstein, 1999). Dark green: hydrophobic residues; light green: aromatic residues; orange: glycine and proline; red: negatively charged residues and cysteine; dark blue: positively charged residues; light blue: small polar residues; magenta: larger polar residues.

The sequence positions of the UCP TM helices used throughout this work are those determined by Aquila *et al.* (1985) using hydrophobicity analyses. While improved methods for TM helix location are likely to have been developed more recently, the high consistency between the positions of aromatic residues shown in Figure 2.9, the predictions made by the hydropathy plot in Figure 2.3, and the positions determined by Aquila *et al.* (1985) suggest that the latter are relatively accurate.

The triplet KVR is found at residue 37 in the first domain and in the corresponding position of the hydrophobic loops in the second and third domains. Interestingly, the only UCPs to lack this exact triplet in their first domains are bovine UCP1 (RYK) and UCP2 (GPR), which also lack several other features that are otherwise characteristic of the UCPs. These substitutions do, however, tend to retain other positive residues at positions 38 and 40, suggesting that their presence is important for UCP function. The cress UCP2 sequence, the most evolutionarily distanced from the other species studied, contains the KVR triplet at position 38 and is the only one in which the motif has shifted.

Deletion of residues 261-269 from UCP1 abolishes nucleotide regulation of transport, suggesting that the loop may form a hydrophobic pocket for binding of nucleotides (Gonzalez-Barroso *et al.*, 1997, 1999). (These studies are described in more detail in Section 2.1.6.3). Similar mutations in the corresponding regions of the first and second domains have given very similar results, suggesting that all three domains are involved (Gonzalez-Barroso *et al.*, 1999).

Further work is needed to experimentally confirm a functional role for the conserved residues and to determine whether the differences in sequence characteristics between UCP1, 2 and 3 represent functional differences between the proteins. The presence of charged residues within the transmembrane helices, regions where hydrophobic residues are generally strongly favoured, suggests that their presence is important for the function of the UCPs. Mutagenesis studies can be used to identify a role for these residues. In addition, an understanding of the 3-dimensional structure and arrangement of the UCP protein is an extremely desirable goal, in order to be able to select the most likely mechanism of transport and identify the groups involved. It is hoped that this work may help to provide some of this much needed structural information concerning the UCPs by establishing a framework for experimental testing.

2.1.7 Mechanism of proton transport

It is not known by what mechanism protons are transported by the UCPs but two contrasting hypotheses have been put forward. In the first, suggested by Klingenberg & Echtay (2001), protons themselves are passed through the channel, down a chain of fatty

acid head-groups thought to act as cofactors. However, according to Skulachev's theory, fatty acid anions are the transported species and the net transfer of protons occurs because protonated fatty acids are able to freely flip-flop across the membrane (Skulachev, 1991).

It has been shown that UCP1 contains a histidine pair that is essential for proton transport but not for Cl^- transport or nucleotide inhibition (Bienengraeber *et al.*, 1998). In support of Klingenberg's theory, the pair is proposed to act as the first pair of proton acceptors/donors in the chain, at the entrance to the proton transport pathway. One of these histidine residues is absent in UCP3 and both are absent in UCP2. Hence, if this histidine pair is involved in proton transport, the UCP1 homologues are likely to use different mechanisms for transport to UCP1.

Evidence for the fatty acid cycling mechanism is now accumulating from work by Garlid *et al.* (1996, 2000). The mechanism requires that (1) the UCPs are only capable of transporting unprotonated fatty acid ions; and (2) protonated fatty acids are capable of flip-flop across the membrane at rates equivalent to that of UCP proton transport. Garlid's group have demonstrated the former (Jezek *et al.*, 1997a), whilst the latter has previously been shown (Kamp *et al.*, 1995). Garlid's group have also identified fatty acids that are incapable of flip-flop across the bilayer (Jezek *et al.*, 1997b). In support of the mechanism, they have shown that these 'inactive' fatty acids are incapable of stimulating UCP-mediated proton transport (Jezek *et al.*, 1997a). Similarly, whilst two sulphonates investigated by Garlid *et al.* (1996) inhibited UCP-mediated Cl^- transport, only the one incapable of membrane flip-flop also inhibited H^+ transport.

Whatever the mechanism of proton transport, it seems clear that a channel is found through the core of the protein, acting as the translocation pathway. Strong evidence for this was provided by Gonzalez-Barroso *et al.* (1999), who showed that the kinetics of UCP1 were converted from that of a carrier to a channel when 'gating' loops were removed from the structures. The UCPs have been termed 'gated pores', like many other carriers and channels showing characteristics of both types of transporter (Arechaga *et al.*, 2001).

It has been suggested that carriers and channels may both require a membrane-spanning channel, typically formed by a bundle of 12 transmembrane helices. This may form the basic building block of both types of transporter, the distinction lying in the accessibility of the binding centre, halfway across the membrane (Arechaga *et al.*, 2001; Jones & George, 2000). If the binding centre is accessible from both sides of the membrane the protein will act as a channel. Alternatively, if, when a substrate enters through a single open entrance pathway, its binding causes a conformational change which closes the entrance gate and opens an exit gate, the protein may be considered a carrier. Removal of the exit gate loops of UCP1 would then enable protons to enter from both sides

of the membrane, permitting the carrier to act as a channel.

2.2 Hypothetical models of the uncoupling proteins, based on the literature

2.2.1 Overview of the experimental evidence for the UCPs as dimeric, single channel proteins

The UCPs consist of relatively hydrophobic α-helices which span the membrane, connected by more polar loops. They are thought to function as dimers. Evidence for the dimeric nature of the mitochondrial carriers has been obtained both for UCP1 and for other members of the family. Cross-linking studies with UCP1 (Klingenberg & Appel, 1989) and the oxoglutarate carrier (Bisaccia *et al.*, 1996) have indicated that the degree of cross-linking is independent of carrier concentration and is observed for both the isolated and native mitochondrial proteins. Molecular weight measurements using both chromatography and electrophoresis have indicated that the molecular weight of UCP1 (Lin *et al.*, 1980) and the citrate carrier (Kotaria *et al.*, 1999), amongst others, is approximately 64 kDa, roughly equivalent to that of a dimer. Schroers *et al.* (1998) showed that heterodimeric constructs of active and inactive forms of the phosphate carrier are totally inactive, suggesting that cross-talk between two monomers is required for function. Finally, expression studies have shown that a covalent tandem dimer of the adenine nucleotide carrier shows native binding and transport properties (Trezeguet *et al.*, 2000). A single study has suggested a tetrameric state for the adenine nucleotide carrier, since only one ATP binding site was detected per four carrier monomers (Dupont *et al.*, 1982), although the work in an accompanying paper supported a dimeric form (Brandolin *et al.*, 1982).

Hence, throughout this study, the UCPs are generally assumed to function as dimers although several monomeric models are also included for comparison. In a dimer, up to 12 helices may be involved in proton transport. How many of them contribute, and their relative arrangements and orientations, however, remain to be established.

As described in Section 2.1.6.1, the UCPs have three strongly homologous domains, each containing two transmembrane-spanning α-helices. The similarity between the domains suggests that an ancestral 2 transmembrane helix protein has duplicated twice to form a trimer, in order to provide the 6 transmembrane helix protein seen today. This suggests that each domain of the UCPs will form a symmetric unit, giving rise to pseudo-3-fold symmetry of the UCP monomer. At present, none of the TM proteins with known

3-dimensional structures contain homologous domains, making it impossible to confirm that sequence homology in membrane proteins leads to structural homology and therefore symmetry. However, based on our current understanding of protein evolution this seems a valid assumption and is virtually always observed in water-soluble proteins. In addition, a number of the TM protein structures analysed in Chapter 3 consist of homologous chains, which would be expected to behave similarly to homologous domains, and these pack to give structural symmetry. Finally, in support of a pseudo-3-fold symmetrical structure and the equivalence of all domains, all three loops on the matrix side of the UCPs are known to contribute to pore gating (Gonzalez-Barroso *et al.*, 1999). As a result of these lines of evidence, it has been assumed throughout this work that each UCP monomer will show pseudo-3-fold symmetry of structure. The structure of the adenine nucleotide carrier indeed confirms this assumption.

It is further assumed that the UCPs contain a relatively large transport pore, despite that fact that some other proteins that transport protons, such as bacteriorhodopsin, do not contain an obvious pore. The assumption that the UCPs have a pore is based upon the knowledge that other members of the family, which are likely to show a very similar structure, transport large molecules like ATP, for which a pore would be required. In addition, Gonzalez-Barroso *et al.* (1997) have shown that if it's gating loops are removed, UCP function will a non-specific pore. Now that the structure of the adenine nucleotide carrier has been determined, this assumption has been validated.

The UCP dimer is thought to contain a single channel for transport, rather than each monomer being capable of transport independently. This is based on the finding, using the related ADP/ATP translocase, that one substrate molecule is transported per dimer, per cycle (Huang *et al.*, 2001). This result is consistent with either the presence of a single channel per dimer or anti-cooperativity of two channels, permitting only one to be open at any time. However, a chimera consisting of one active and one inactive monomer is totally inactive (Huang *et al.*, 2001). A two channel model would predict 50% activity of this chimera. This therefore suggests that each dimer contains a single transport pore, for which both monomers are required to be functional. Previously, a similar result was found for the phosphate transporter (Schroers *et al.*, 1998). The high sequence similarity amongst the mitochondrial carrier proteins suggests that they will share these general topological and functional characteristics. Hence most of the models considered here contain a single TM channel. However, since this evidence is relatively weak, one two-channel model was included for comparison.

After completion of this work the structure of another member of the MCF, the adenine nucleotide carrier, was determined by Pebay-Peyroula *et al.* (2003). The structure, described in detail in Chapter 4, Section 4.4.1, shows that the protein has six TM he-

lices with pseudo-3-fold symmetry as predicted. However, the structure consists of a monomeric unit containing a single transport pore. It is therefore unclear whether the predicted dimeric nature of the family is biologically correct. As discussed later, these incorrect assumptions are likely to have had an impact on the accuracy of the models generated in this chapter.

2.2.2 Information from mutagenesis studies

Information gained from experimental work can be used to impose constraints on UCP helix organisation, allowing us to select between different possible arrangements. Mutagenesis studies, for example, which indicate that specific residues are involved in transport, will enable particular faces of some helices to be assigned a pore-lining function. D27, for example, is located on the first transmembrane helix and is conserved throughout all uncoupling proteins, but not found in other members of the mitochondrial carrier family (Klingenberg & Echtay, 2001). H^+ transport is abolished when D27 is mutated to N, but not when its negative charge is maintained by mutation to E. This suggests that helix 1, and in particular its negative aspartate, play a crucial role in transport, and therefore may be involved in pore formation. However, a direct role for this residue in proton transport has been disputed, and it was suggested that the D27 negative charge may be essential for formation of a competent conformation, via the formation of a hydrogen bond to E190 (Hagen & Lowell, 2000).

The conserved arginines (R83, R182 and R276) are each located at homologous positions in relatively conserved faces of helices 2, 4 and 6. They have been shown to be required for binding of nucleotides to UCPs (Echtay *et al.*, 2001a). On the one hand, since the related ADP/ATP translocase transports nucleotides, this may suggest that the nucleotide-binding site of the UCPs will be within its transport channel and that R83, R182 and R276 may have a pore-lining role. This of course assumes that the transport pathway is conserved between the two proteins, (and in fact all other proteins in the family), despite their diverse substrates. In support of this hypothesis, mutations in the ADP/ATP translocase, in the residues corresponding to R83, R182 and R276, prevent translocation (Muller *et al.*, 1996). A pore-lining role for these arginines would provide support for a model of UCP transmembrane structure in which helices 2, 4 and 6 all contribute to the pore.

However, whilst mutation of R83, R182 and R276 to neutral residues prevents binding of nucleotides to UCP1, proton transport in the mutants is unaffected (Echtay *et al.*, 2001a). There are two possible explanations for these results: (1) that nucleotide-binding and proton transport occur at different sites and the conserved arginines do not line the

pore, or (2) that a common channel is used but different residues are involved.

Finally, the conservation of these residues throughout members of the mitochondrial carrier protein family that do not bind nucleotides suggests they do not play a direct role in nucleotide-binding. Instead, their role may be in maintenance of the native conformation, via critical contacts with other residues. Hence the finding that the homologous arginines are essential for nucleotide-binding does not necessarily constrain them to a pore-lining location.

These examples therefore illustrate the difficulty in using this type of information in model prediction. Further studies are urgently needed to resolve these issues so that information regarding the role of these residues can be used with more confidence in UCP helix model selection.

2.2.3 Potential models for transmembrane arrangements of the UCPs

The models developed from the above information are illustrated in Figure 2.10. Whilst they are not consistent with all of the experimental data in the literature, they all conform to the assumptions outlined above of a pore-containing dimeric form with pseudo-3-fold symmetry. Several variations upon Model 1 exist, which are shown in Figure 2.11. Similarly for Model 3, any sequential pair of helices could be found at the monomer-monomer interface. The positions of TM helices, illustrated in figure 2.9, are taken from (Aquila *et al.*, 1985), who first cloned the UCP1 sequence.

An assumption made in the generation of UCP TM helix models is that folding occurs sequentially, so that in the structure each helix is located adjacent to the next helix in the sequence. Maintaining pseudo-3-fold symmetry requires that in each sequence repeat the pair of helices shows the same arrangement relative to one another. This greatly reduces the number of possible models with non-sequential folding to one alternative per model. It is impossible to distinguish between each sequential model and its non-sequential alternative using the current method, since each helix must occupy an equivalent position and it is simply the identities of its neighbours that change. Therefore, if one model is selected to represent the actual UCP structure, its non-sequential alternative is also a possible candidate. Cross-linking studies would be needed to distinguish between them by determining the helix neighbour relationships.

Since Bowie (1997) showed that 37 out of 38 TM helices in his dataset were found adjacent to at least one sequence neighbour, the sequentially numbered models appear most likely. Unfortunately, however, the recently published structure of the lactose permease (Abramson *et al.*, 2003) does not show this arrangement so clearly, illustrating that the

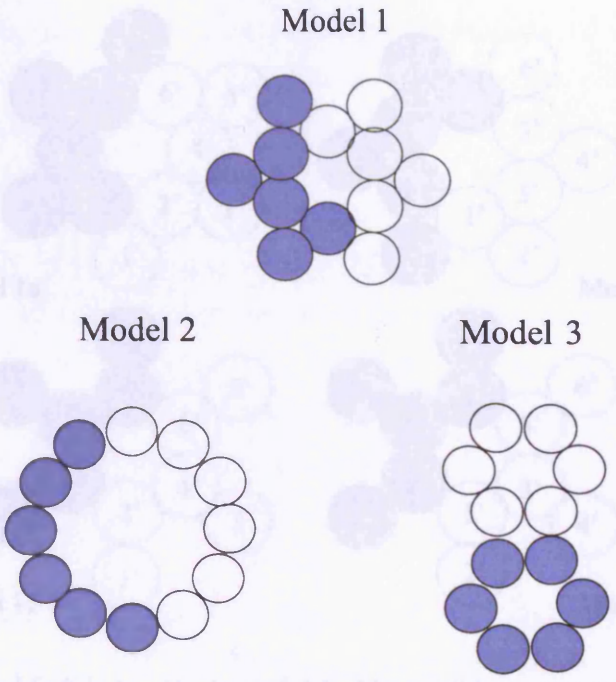


Figure 2.10: Models 1, 2 and 3: Possible arrangements of uncoupling protein transmembrane helices. Each helix, represented as a circle, has a diameter of 10Å Each monomeric unit has 6 TM helices, and one is shaded per dimer. Model 1: 6 of the 12 dimer helices contribute to the pore, producing a pore with a diameter of approximately 10Å. Model 2: All of the 12 dimer helices contribute to the pore, producing a pore with a diameter of approximately 30Å. Model 3: Each monomer forms a separate transport channel with a diameter of approximately 10Å.

non-sequential models can not be entirely excluded. The following sections describe the relative merits of each model.

2.2.3.1 Model 1

All forms of Model 1 (Figure 2.11) show the predicted pseudo-3-fold symmetry. However, Models 1a and 1c, in which helices 2, 4 and 6 surround the pore, do not permit a pore-lining location for D27 on helix 1, contrary to the evidence described in Section 2.2.2. Hence, if Model 1a or 1c is correct, we must assume that D27 contributes only indirectly to transport. Conversely, Models 1b and 1d do not permit the pore-lining location of the homologous arginines on helices 2, 4 and 6, and hence assume that these residues participate in nucleotide-binding indirectly.

If D27 or the homologous arginines are not found in a pore-lining position, they must

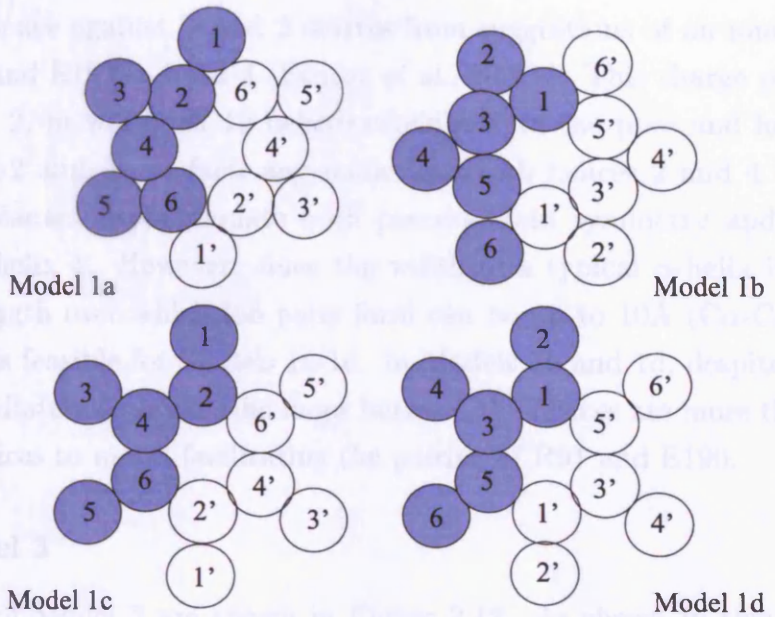


Figure 2.11: Models 1a, 1b, 1c and 1d: Alternative arrangements of uncoupling protein transmembrane helices for Model 1. Each helix, represented as a circle, has a diameter of 10Å.

be either accessible to membrane lipid-tails or buried unpaired within the protein core, since, as cen be seen in Figure 2.9, no likely pairing charge is available in the other helices. Whilst intuitively both of these alternatives may seem unlikely, later work in the current study (Chapter 3, Section 3.3.8) suggests that there are many reasons why charged residues may be accessible to membrane lipid-tails. In addition, it has been shown that charged residues will often hydrogen bond with polar residues or phospholipid headgroups, so that a pairing charge is not required (Adamian & Liang, 2001; this study). As a result, neither knowledge of TM protein structure or experimental evidence constrains the location of any of these residues to the pore or any other location. Hence, none of these models can be excluded on the basis of the position of these residues.

2.2.3.2 Model 2

Model 2 shows pseudo-3-fold symmetry and allows D27, and the homologous arginines R83, R182 and R276 to line the pore. Model 2 requires that a gate is present to restrict transport through the very large pore to that of protons. There is evidence in support of this, as described in Section 2.1.7. However, the pore in Model 2 is far greater in size than that required for the transport of protons, fatty acids or ATP, and the benefits of

the evolution of such a large pore are not, therefore, apparent.

Further evidence against Model 2 derives from suggestions of an ionic bond between R91 in helix 2 and E190 in helix 4 (Echtay *et al.*, 2001a). This charge pairing could not occur in Model 2, in which all 12 helices contribute to the pore and helix 3 is located between helices 2 and 4. In fact, any model in which helices 2 and 4 are adjacent for charge pairing cannot accommodate both pseudo-3-fold symmetry and the pore-lining role of D27 in helix 1. However, since the width of a typical α-helix is approximately 10Å and the length over which ion pairs form can be up to 10Å (Cα-Cα distance), the R91-E190 pair is feasible for Models 1a-1d. In Models 1b and 1d, despite helices 2 and 4 not being immediately adjacent, the loops between the helices are more than long enough to allow the helices to move, facilitating the pairing of R91 and E190.

2.2.3.3 Model 3

The variations of Model 3 are shown in Figure 2.12. As shown in this figure, Model 3 includes 3 monomeric models (Models 3s, 3t and 3u), in addition to 18 dimeric variations that can be grouped into 3 classes: those in which the even and odd-numbered helices occupy equivalent positions, (Models 3a-f) those in which TMs 2,4 and 6 are more peripheral (Models 3g-l) and those in which TM2 2, 4 and 6 are more peripheral (Models 3g-l) and those in which TMs 1, 3 and 5 are more peripheral (Models 3m-r).

The model shows pseudo-3-fold symmetry of each monomer and permits a pore-lining role for D27 and the conserved homologous arginines. It is, however, inconsistent with an ionic bond between helices 2 and 4. Models 3s, t and u are contrary to the evidence that suggests that the UCPs are functional as a dimer. All variations of Model 3 contain an equivalent of two transport channels per dimer. Whilst there is evidence to support mainly a single channel model (Schroers *et al.*, 1998; Huang *et al.*, 2001), it was felt that this evidence was not sufficiently strong to permit this model to be entirely discounted.

2.2.4 Geometric Considerations

Klingenberg & Appel (1989) have shown that the cross-linking of UCP monomers into a dimer can occur between the C304 residues, at the C terminal end of each protein. Homodimerisation of two identical molecules of UCP will always result in the juxtaposition of each N terminus with the C terminus of the other monomer, as shown in all models. It seems extremely unlikely that two different 'handed' isoforms of UCP heterodimerise to allow close approach of their C termini. Hence the 15 residues of extended chain from the end of helix 6 to the cross-linking cysteine (residues 289 to 304) must be sufficiently long to reach across the pore to C304 of the other monomer. Assuming a length of 3.4Å

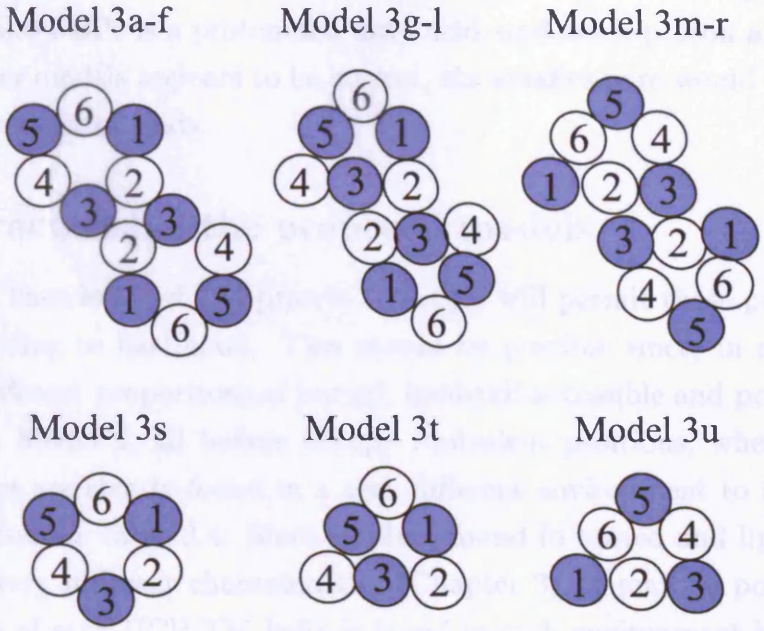


Figure 2.12: Models 3a-u: Alternative arrangements of uncoupling protein transmembrane helices. Models 3b, 3h and 3n are shown as representative examples of each class of dimeric models, although any 2 sequential helices may form the monomer-monomer interface for each class. Each helix, represented as a circle, has a diameter of 10Å. TMs 1, 3 and 5 are shaded to highlight their different positions in each group of models. Model 3s-u are a monomeric forms of Models 3a-f.

per residue for extended chain, this imposes the constraint that helix 6 of each monomer may not be more than 100Å apart. Considering the predicted dimensions of the models described, however, this would be unlikely to impose constraints upon the packing of the other helices.

The close sequence similarity of the UCPs to the ADP/ATP translocase suggests that they may have retained a pore wide enough for the transport of ATP. That they are regulated by the binding of nucleotides further supports this idea. All of the models considered here have pores large enough for the transport of ATP (the estimated size of an extended molecule of ATP is 6 by 15Å).

The arrangement of UCP TM helices has obvious implications for the mechanism of proton transport. Model 2, for example, creates a pore with an area equivalent to that of more than 6 alpha helices or approximately 480Å^2, far greater than that thought to be required for the transport of a lone proton. It is very difficult to suggest how residues from all helices may interact with a proton of extremely small size passing through the centre.

If Model 2 appears to be correct, this will lend support to the theory that the species transported by the UCPs is a protonated fatty acid, and not a proton alone. Conversely, if one of the other models appears to be correct, the smaller pore would permit transport of either protons or fatty acids.

2.2.5 Characterising the proposed models

It is hoped that knowledge of TM protein structure will permit these possible models to be ranked according to likelihood. This should be possible since, in each model, each TM helix has different proportions of buried, lipid-tail-accessible and pore-lining surface. For example, in Model 2, all helices occupy equivalent positions, whereas in Model 3, two of the helices are clearly found in a very different environment to the other helices. This is summarised in Table 2.4. Since residues found in buried and lipid-tail-accessible positions show very different characteristics (Chapter 3), it may be possible to identify what proportion of each UCP TM helix is found in each environment by analysis of the sequence. However, it has not been possible to identify significant differences between pore-lining and buried residues, (Chapter 3) and therefore any method will need to be based on the discrimination of buried and lipid-tail-accessible residues alone.

Model	Helices	Proportion Buried (%)	Proportion Accessible (%)	Proportion in Pore (%)
1a	2 4 6	66	17	17
	1 3 5	33	67	0
1b	1 3 5	66	17	17
	2 4 6	33	67	0
1c	1 3 5	17	83	0
	2 4 6	50	33	17
1d	2 4 6	17	83	0
	1 3 5	50	33	17
2	1 2 3 4 5 6	33	42	25
3a-r	1 2	50	33	17
	3 4 5 6	33	50	17
3s-u	1 2 3 4 5 6	37	46	17

Table 2.4: Estimated percentages of helix surfaces found buried in the protein, accessible to lipid-tails or lining the pore for each of the UCP models. Percentages were estimated using trigonometry, assuming each helix to be perfectly circular and non-overlapping, as shown in the models in Figure 2.10. Other TM helices could be located at the monomer-monomer interface in Model 3. Only TMs 1 and 2 are shown here for brevity.

2.3 Conclusions

In conclusion, according to the current data in the literature, there is no single helical arrangement that is consistent with all of the available experimental data. A summary of which models are consistent with which experimental data is given in Table 2.5. Most models are supported by 6 out of 7 pieces of data, making it impossible to definitively either support or refute any of them. The problem lies in the difficulty in interpreting the results from mutational experiments: it is often impossible to determine whether the loss of function that occurs on mutation is caused directly by the loss of a participating residue, or indirectly due to disruption of the native conformation. Hence, much of the experimental evidence described here is not definitive enough to use for model prediction in this context. It does, however, seem that Models 1a-1d or 3 may be the most likely, based on their pseudo-3-fold symmetry and small pore size.

Evidence	**1a/c**	**1b/d**	**2**	**3**
Single pore	Y	Y	Y	N
3-fold symmetry	Y	Y	Y	Y
Pore large enough for transport of ATP	Y	Y	Y	Y
D27 in helix 1 lines pore	N	Y	Y	Y
Ionic bond between helices 2 and 4	Y	Y	N	N
Pore lining role for conserved arginines	Y	N	Y	Y
Helix 6 of each monomer within 100Å	Y	Y	Y	Y
Total	6	6	6	5

Table 2.5: A summary of which models are consistent with various pieces of experimental data. Y indicates that the model is supported by the evidence, N indicates that it is not. The total number of pieces of supporting evidence is also given for each model. See Section 2.2.3 for references and discussion of this data.

The structure of the adenine nucleotide carrier, a protein related to the UCPs, was determined after this work was completed (Pebay-Peyroula *et al.*, 2003). The structure is described in detail in Chapter 4, Section 4.4.1. Consistent with the assumptions made,

the structure shows that the monomer has six TM helices surrounding a pore with pseudo-3-fold symmetry. However, the structure consists of a monomeric unit and it is therefore unclear whether the assumption made here of a dimeric structure for the family is biologically correct. As described in Chapter 4, the adenine nucleotide carrier is consistent with the majority of the experimental evidence used to select between models in this chapter, indicating that the data used is in most cases reliable. It seems that experimental data of this type is, at least in the case of the UCPs, largely consistent with a considerable number of alternative models, in addition to the correct one. This suggests that the inability to identify a single correct model for the UCPs in this chapter is due to the lack of discriminatory power that such data have during modelling.

Two general approaches can be suggested for the prediction of a model for the arrangement of the UCP TM helices. The first is manual model building based on experimental data from the literature, as performed here. The results suggest that the method is unlikely to be effective, unless experimental data with considerably more discriminatory power become available. The second method is based on prediction using the general principles of TM protein structure, and requires the greatest understanding of these features that can be obtained. This approach, and the necessary analysis of TM protein structure, will be followed in subsequent chapters.

Chapter 3

Computational analysis of transmembrane protein structure

3.1 Introduction

3.1.1 Transmembrane protein structure

Transmembrane (TM) proteins are those that span the membrane lipid bilayer. They are estimated to comprise 20-30% of all proteins (Arkin *et al.*, 1997; Wallin & von Heijne, 1998; Schwartz *et al.*, 2001; Knight *et al.*, 2004; Klein *et al.*, 2004) and are of huge biological significance since they mediate most of the communication between cells and cellular compartments.

The majority of TM proteins consist of relatively hydrophobic α-helices, which span the membrane, connected by more polar loops. These are known as the α-bundle TM proteins. In contrast, the β-barrel family of TM proteins span the membrane via β-sheets and are not considered in this analysis. As shown in Figure 3.1, the central region of the membrane consists of a highly hydrophobic 30Å thick lipid-tail region, formed by the hydrocarbon chains of the phospholipids. This is surrounded on either side by a 15Å thick region formed by the highly polar phospholipid head groups. These regions of the TM helices, will be termed the 'lipid-tail-spanning' and 'head-group-spanning' regions respectively, throughout this work. In these two environments the transmembrane helices will be subject to highly differing constraints, which are likely to be manifest in differences in amino acid content.

While Figure 3.1 illustrates the membrane as a very static and ordered structure, this is in fact a considerable over-simplification. Water is not only present on either side of the membrane, but is also abundant within the outer edges of the head-group region. In contrast, water is almost entirely excluded from the highly hydrophobic core of the

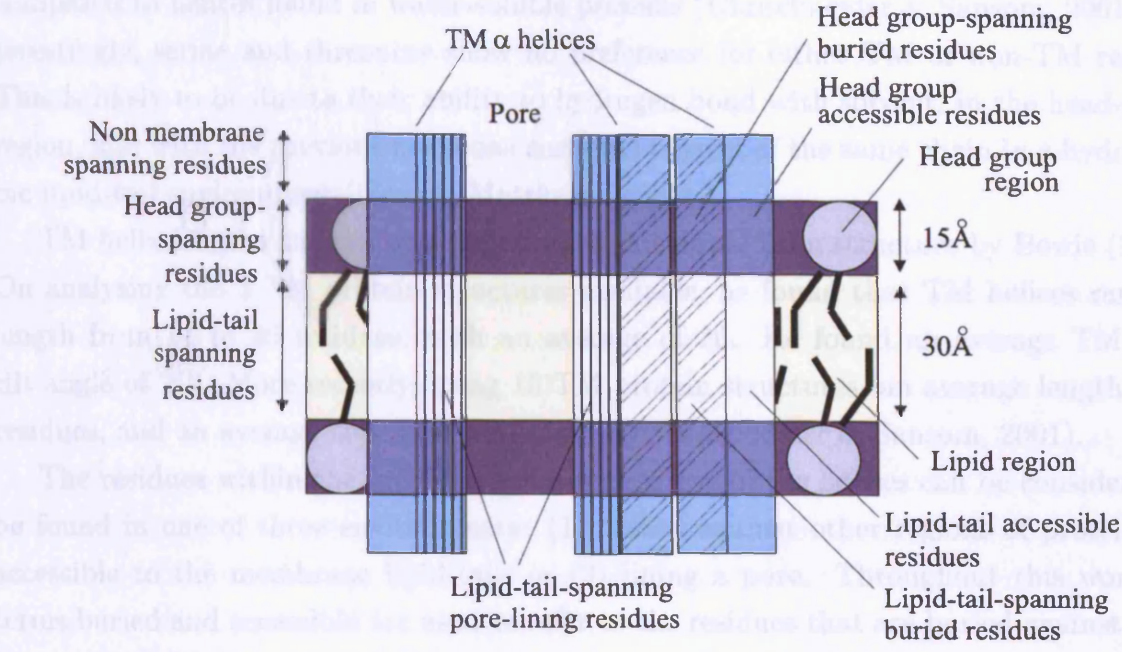


Figure 3.1: Schematic diagram showing the structure and thickness of a typical membrane.

membrane. A gradient of water density exists through the membrane, between these two extremes. The membrane is fluid, and phospholipids constantly more around in the plane of the membrane, interacting with each other and other molecules, such as proteins, that are found in the membrane.

In addition, the membrane contains a mixture of different phospholipids with varying chemical properties and tail lengths. Different organisms, particularly when comparing eukaryotes and bacteria, have very different membrane phospholipid compositions, leading to considerable variation in membrane properties such as thickness (Bretscher & Munro, 1993; Killian, 1998; Williamson *et al.*, 2003; O'Keeffe *et al.*, 2000). Hydrophobic mismatch is said to occur when the length of the hydrophobic membrane-spanning segments of a TM protein differ from the thickness of the hydrophobic region of the membrane. The thickness of the membrane, and the conformation of embedded TM proteins, are thought to adjust to minimise hydrophobic mismatch (Killian, 1998; Harroun *et al.*, 1999; Liu *et al.*, 2004a,b). This may be achieved by alterations in the conformation of phospholipid tails or in the tilt of TM helices. As a result, the conformation, and hence the function, of a TM protein may depend critically upon the composition of the membrane in which it is found (De Planque *et al.*, 2004; Wiggins & Phillips, 2004).

From analysis of 15 TM protein structures, it was observed that TM helices are enriched in hydrophobic amino acids and they contain fewer charged and polar amino acids,

compared to helices found in water-soluble proteins (Ulmschneider & Sansom, 2001). Interestingly, serine and threonine show no preference for either TM or non-TM regions. This is likely to be due to their ability to hydrogen bond with solvent, in the head-group region, and with the previous backbone carbonyl oxygen of the same chain in a hydrophobic lipid-tail environment (Gray & Matthews, 1984).

TM helix lengths and packing angles were calculated from structure by Bowie (1997). On analysing the 3 TM protein structures available, he found that TM helices range in length from 14 to 36 residues, with an average of 26. He found an average TM helix tilt angle of 21°. More recently, using 15 TM protein structures, an average length of 27 residues, and an average tilt of 22° were found (Ulmschneider & Sansom, 2001).

The residues within the lipid-tail-spanning regions of the helices can be considered to be found in one of three environments: (1) buried against other regions of protein, (2) accessible to the membrane lipid-tails or (3) lining a pore. Throughout this work the terms buried and accessible are used to refer to the residues that are buried against other parts of the TM helix bundle or accessible to membrane lipid-tails respectively, rather than the more common usage of buried or accessible relative to water.

The different environments of lipid-tail-accessible and buried residues result in different sequence characteristics. Hence it can be postulated that lipid-tail-accessible residues will more often be hydrophobic in character than will buried residues, since the protein must be stable when folded in the fatty, hydrophobic membrane. Early analysis of the preferences of residues for buried or lipid-tail-accessible positions was performed on the photosynthetic reaction centre of *Rhodobacter sphaeroides* (Rees *et al.*, 1989). Whereas, in this TM protein, the lipid-tail-accessible residues were more hydrophobic than those buried, (Rees *et al.*, 1989) in water-soluble proteins the buried residues were more hydrophobic (Chothia, 1976), reflecting the different environments in which these classes of protein are found.

The buried residues in the reaction centre had a very similar hydrophobicity to residues buried within water-soluble proteins, as determined previously by Chothia (1976). Hence, water-soluble and TM proteins could be broadly considered to be internally stabilised in similar ways, with surface residues modified in ordered to facilitate solubility in the required medium. This finding confirms the results of other groups (Stevens & Arkin, 1999, 2000; Rees & Eisenberg, 2000), that TM proteins are not simply 'inside-out' water-soluble proteins, with highly polar internal regions, as has been proposed (Engelman & Zaccai, 1980). It is important to confirm these results using the new TM protein structures that are now available.

It is likely that lipid-tail-accessible residues will be less conserved in terms of their sequence than buried residues. This is because there will be strong selective pressure to conserve buried residues, so that they continue to interact favourably with their interac-

tion partners on other helices. There will be much less selective pressure for lipid-tail-accessible residues to be conserved, since they are not involved in structurally important contacts with other protein regions. Mutations in lipid-tail-accessible residues are therefore much less likely to disrupt the folding, and hence the function, of the protein. As a result, throughout evolution, those individuals in which a protein contains a mutation in a buried TM residue are likely to be less viable than those in which the protein contains only mutations in lipid-tail-accessible residues. Hence, mutations in buried residues are less likely to be inherited and sequence variation between homologues will be much less at buried positions than lipid-tail-accessible. In support of this hypothesis, the buried residues have been shown to be more conserved than the lipid-tail-accessible ones in the photosynthetic reaction centre of *Rhodobacter sphaeroides* (Rees *et al.*, 1989) and more recently in a group of TM protein structures (Stevens & Arkin, 2001). The present study determines the degree to which these trends of hydrophobicity and conservation hold for a much larger dataset of TM proteins.

3.1.2 Packing of TM helices

In water-soluble proteins, the helices of left-handed coiled-coils interact using a mechanism known as 'knobs into holes' or leucine zipper packing (Crick, 1953). This has also been observed to be a common motif in TM helices (Langosch & Heringa, 1998). The motif consists of a heptad repeat (*abcdefg*), with hydrophobic residues found at the positions buried between the helices, *a*, *d*, *e* and *g*. The residues at position *d* tend to be leucines and at position *a* tend to be valines. The side chains of these residues interact via extensive hydrophobic interactions, forming alternate leucine and valine 'rungs' connecting the two helix backbones.

Despite the similar hydrophobicity of buried residues in water-soluble and TM proteins, and the use of leucine zipper packing, there are several differences in the way that their helices pack together. TM helices associate more tightly than the helices in water-soluble proteins (Eilers *et al.*, 2000). This may be due to the fact that the residues buried between TM helices tend to have shorter side chains than those that are lipid-tail-accessible (Jiang & Vakser, 2000), allowing closer approach of the helix backbones. In support of this, several groups have identified an inverse relationship between the residue volume and its tendency to be found buried at a helix interface (Javadpour *et al.*, 1999; Eilers *et al.*, 2000; Adamian & Liang, 2001). The most common residues to be buried between TM helices are glycine, alanine, serine and threonine. In contrast, leucine and alanine are the most frequently buried residues in water-soluble proteins. Eilers *et al.* (2000) suggested that perhaps this close approach, via small polar residues is necessary

to compensate for the lack of the hydrophobic effect as a driving force for the folding of TM proteins.

Javadpour *et al.* (1999) found that glycine, alanine, serine, threonine, cysteine and proline are preferred in buried positions in TM proteins, whereas leucine, isoleucine, glutamine and, unexpectedly, the charged residues prefer lipid-tail-accessible positions. This work involved only 4 TM protein structures, and there is therefore a need to repeat the calculations for the new structures available. This need is highlighted by the fact that a later study involving 15 structures by Ulmschneider & Sansom (2001) found that leucine, isoleucine, valine and phenylalanine showed no preference for either buried or lipid-tail-accessible positions. The latter authors did, however detect a significant preference of glycine and alanine for buried positions, in agreement with the earlier work. The lack of significant discrimination for the hydrophobic residues may be due to the use of a relatively large accessible surface area cut-off (10%) to select buried and accessible residues. This may have increased the pool of residues classified as buried to include a significant percentage that are actually partially lipid-tail-accessible. In addition, the results of this study are likely to have been biased by the inclusion of multiple members of 4 of the protein families in the dataset.

A striking trend is the strong preference of glycine for buried positions (Javadpour *et al.*, 1999; Eilers *et al.*, 2000; Adamian & Liang, 2001; Ulmschneider & Sansom, 2001). The role of glycine has been specifically studied by Javadpour *et al.* (1999). With its side chain consisting of a single hydrogen atom, glycine allows adjacent helices to approach more closely than any other residue. Perhaps as a result of this, in cytochrome C oxidase, it is often found at helix crossings where, Javadpour *et al.* (1999) have suggested, it may function as a 'molecular notch' for orientating one helix against another.

The presence of a glycine residue also exposes the polar backbone atoms of its own chain, facilitating the formation of hydrogen bonds and dipolar interactions (Javadpour *et al.*, 1999). These forces are likely to be particularly strong in the hydrophobic environment of the bilayer and hence may play a major role in protein stability. Whilst interactions between the backbone atoms of different helices are rare in both classes of protein, they seem to be more common in TM proteins than in water-soluble proteins. In fact, whilst interactions of the backbone account for only 1.1% of all atomic contacts in water-soluble proteins, they comprise 2.8% of the contacts in TM proteins (Adamian & Liang, 2001). Interactions with the backbone are most common where glycine residues interact.

3.1.3 Common residue interactions at helix interfaces

Adamian & Liang (2001) performed an analysis of the frequency with which particular pairs of interacting residues on adjacent helices were observed. They found that some pairs occurred at much higher frequency than would be predicted from the residue occurrences by chance alone. These included basic residues interacting with aromatic residues, particularly W-R, W-H and Y-K. This is presumably due to an interaction between the cation and the pi electron cloud, as has been suggested by White (2001) for tryptophan.

In water-soluble proteins, helix-packing is mainly mediated by the interaction of pairs of hydrophobic residues. In addition, both water-soluble and TM proteins contain pairs of charged residues, forming salt bridges between the helices. Disulphide bonds appear to be far less common in TM proteins than water-soluble proteins (Adamian & Liang, 2001). However, one major difference between the packing of the two classes of proteins is the presence, in TM proteins alone, of many inter-helix interactions between two polar residues, or between a polar and a charged residue. Hence, charged residues in the lipid-tail environment need not be paired with an opposing charge, and their hydrogen bonding requirements are often met by polar residues, particularly asparagine and glutamine. Adamian & Liang (2001) described the helix-packing interactions of TM proteins as involving a far more diverse set of polar residues than those in water-soluble proteins. Whilst, in water-soluble proteins, they observed 3 types of interaction between polar groups on interacting helices, in TM proteins they found 22 different types of polar interactions. Hence, in addition to the leucine zipper packing using hydrophobic residues found in both classes of protein, TM proteins also make use of a wide variety of interactions between polar and charged residues. The greater importance of polar interactions in TM proteins is to be expected, due to their greater strength in the low dilectric lipid-tail environment. Similarly, hydrophobic interactions are likely to be weaker between TM helices, and so they would be expected to play a less important role.

3.1.4 Aims of this chapter

Practical difficulties in expressing and crystallising membrane proteins have lead to very few TM protein structures being available for analysis. As a result, due to lack of TM protein structural data, previous work has until recently been based on few membrane protein structures, on datasets biased by the inclusion of multiple homologues, or simply on primary sequence. The importance of the use of data derived from structure and not sequence is highlighted by the work of Ulmschneider & Sansom (2001), who compared the effectiveness of these two data sources in discriminating between TM and non-TM-spanning regions.

Since there has recently been a rapid rise in the number of membrane protein structures solved (more than twice as many are now available as when the last comprehensive study was performed in 2001), it seems likely that it will now be possible to resolve with more confidence previous conflicting results. To our knowledge, the more recent studies analysing a number of structures have not included sequence conservation information. In addition, no studies have analysed whether the observed differences between lipid-tail-accessible and buried residues are sufficient to distinguish between the two environments in a predictive way. These issues are addressed in the current work.

Like the analysis work, mainly due to lack of data TM structure prediction methods have often been based only on sequence data or on very small numbers of TM protein structures (Donnelly *et al.*, 1993; Pilpel *et al.*, 1999). In addition, a number of methods have made use of residue potentials derived from the structures of water-soluble proteins (Fleishman & Ben-Tal, 2002; Pellegrini-Calace *et al.*, 2003; Chen & Chen, 2003). Given the considerable differences between the packing and hydrogen bonding of TM and water-soluble protein helices (Rees *et al.*, 1989; Eilers *et al.*, 2000; Jiang & Vakser, 2000; Javadpour *et al.*, 1999; Ulmschneider & Sansom, 2001; Adamian & Liang, 2001), the use of data derived from water-soluble proteins is likely to limit the accuracy of these predictive methods.

There is therefore a need to identify what features, if any, of TM helix packing would be applicable for use in the prediction of the tertiary structure of TM proteins. In the current work, the amino acid composition of 24 non-homologous α-helical membrane proteins with known structures was analysed. (A non-homologous set is defined here as containing no members sharing more than 20% sequence identity with another member). Only polytopic proteins were selected, (those spanning the membrane more than once), since those with only one TM helix could contain no buried residues within the membrane-spanning region. Sequence conservation, hydrophobicity and amino acid propensities were compared between lipid-tail-accessible and buried residues. Differences between the two groups of residues with the potential for use in TM protein structure prediction are identified and discussed, in order to determine whether prediction of 3-dimensional structural models from protein sequence characteristics alone will be feasible in the near future. Interestingly it is found that several charged and polar residues prefer lipid-tail-accessible positions to buried and that many of the hydrogen bonds made by side-chains are intra-helical.

3.2 Methods

3.2.1 Overview of methods

The residue characteristics of TM helices, derived from structure, were analysed. In order to ensure that the transmembrane segments used in the analysis were as accurately located as possible, the program PSlice was developed to extract the precise position of these from the 3-dimensional coordinates, on the basis of hydrophobicity. This section describes the methods used, which can be divided into three stages. Firstly, the techniques used to generate the dataset of TM proteins for analysis are given. Secondly, PSlice, the algorithm for identification of TM helices from the dataset proteins, is described. Finally, the programs and methods used for analysis of these TM helices are detailed.

3.2.2 Dataset generation

Membrane proteins of known 3-dimensional structure were identified from a website of membrane protein resources (*http://blanco.biomol.uci.edu/MemPro_resources.html*). Proteins with only one TM helix or those that span the membrane with β-sheets were excluded. Cofactors were not removed from the structures to avoid the creation of large cavities within the structures, which may have made the cofactor-binding residues appear to be accessible.

At the time of dataset assembly (January 2004), the structures of 77 α-helical polytopic TM proteins had been determined and deposited in the PDB. These proteins are listed in Table 3.1.

If two of the TM proteins of known structure shared more than 25% sequence identity with another, or were known to be from the same family, only the relative with the highest resolution structure was retained in the dataset. Mutant or partial structures, or those in complex with other molecules such as antibodies, were also excluded, since it was felt that these factors may have affected the native conformation of the protein. This is illustrated in Table 3.1. 53 proteins were removed in this way, leaving 24 non-homologous proteins, listed in Table 3.3.

	PDB	Resolution	Notes	Reference
			Photosystems	
1	*1jb0**	*2.5*		***Jordan* et al.** *(2001)*
2	1vf5	3.0		Kurisu *et al.* (2003)
3	1izl	3.7		Kamiya & Shen (2003)
4	1fe1	3.8		Zouni *et al.* (2001)
5	2pps	4.0		Schubert *et al.* (1997)
			Cytochrome B6f	
6	*1um3*	*3.0*		***Kurisu* et al.** *(2003)*
7	1q90	3.1		Stroebel *et al.* (2003)
			Bacterial Rhodopsins	
8	*1c3w*	*1.6*		***Luecke* et al.** *(1999b)*
9	1e12*	1.8		Kolbe *et al.* (2000)
10	1qhj	1.9		Belrhali *et al.* (1999)
11	1h2s	1.9		Gordeliy *et al.* (2002)
12	1qko/p	2.1		Edman *et al.* (1999)
13	1h68	2.1		Royant *et al.* (2001)
14	1brx	2.3		Luecke *et al.* (1998)
15	1jgj	2.4		Luecke *et al.* (2001)
16	1ap9	2.5		Pebay-Peyroula *et al.* (1997)
17	1brr	2.9		Essen *et al.* (1998)
18	1at9	3.0		Kimura *et al.* (1997b)
19	2brd	3.5		Grigorieff *et al.* (1996)
20	1c8r/s	1.8	Mutant structure	Luecke *et al.* (1999a)
			Protein-Coupled Receptors	
21	*1l9h**	*2.6*		***Okada* et al.** *(2002)*
22	1f88	2.8		Palczewski *et al.* (2000)
			Inward rectifying K^+ channel	
23	*1p7b*	*3.6*		***Kuo* et al.** *(2003)*
24	1n9p	1.8	Partial structure	Nishida & MacKinnon (2002)
			Other K^+ channels	
25	*1bl8**	*3.2*		***Doyle* et al.** *(1998)*
26	1orq	3.2		Jiang *et al.* (2003)
27	1lnq	3.3		Jiang *et al.* (2002)
28	1ors	1.9	Partial structure	Jiang *et al.* (2003)

Table 3.1: *continued*

	PDB	Resolution	Notes	Reference
29	1k4c/d	2.0	Antibody complex	Zhou *et al.* (2001)
			Clc Cl^- channels	
30	***1kpl****	***3.0***		***Dutzler* et al. *(2002)***
31	1kpk	3.5		Dutzler *et al.* (2002)
32	1ots	2.5	Antibody complex	Dutzler *et al.* (2003)
			Acetyl choline receptor pore	
33	1oed	4.0	Partial structure	Miyazawa *et al.* (2003)
			MscL K^+ channel	
34	***1msl****	***3.5***		***Chang* et al. *(1998)***
			MscS K^+ channel	
35	***1mxm****	***3.9***		***Bass* et al. *(2002)***
			Aquaporin/Glycerol facilitator channel	
36	***1j4n***	***2.2***		***Sui* et al. *(2001)***
37	1fx8	2.2		Fu *et al.* (2000)
38	1rc2	2.5		Savage *et al.* (2003)
39	1h6i*	3.5		de Groot *et al.* (2001)
40	1ih5	3.7		Ren *et al.* (2001)
41	1fqy	3.8		Murata *et al.* (2000)
			SecYE protein conducting channel	
42	***1rhz***	***3.5***		***Van den Berg* et al. *(2004)***
			Multi-drug efflux transporters	
43	***1iwg****	***3.5***		***Murakami* et al. *(2002)***
44	1oy6	3.7		Yu *et al.* (2003)
			Major facilitator superfamily transporters	
45	***1pw4***	***3.3***		***Huang* et al. *(2003)***
46	1pv6	3.5		Abramson *et al.* (2003)
			ABC transporters	
47	***1l7v****	***3.2***		***Locher* et al. *(2002)***
48	1pf4	3.8		Chang (2003)
49	1jsq	4.5		Chang & Roth (2001)
			P-type ATPases	
50	***1eul****	***2.6***		***Toyoshima* et al. *(2000)***

Table 3.1: *continued*

	PDB	Resolution	Notes	Reference
51	1su4	2.6		Toyoshima *et al.* (2000)
52	1iwo	3.1		Toyoshima & Nomura (2002)
			Photosynthetic reaction centres	
53	***1eys***	***2.2***		***Nogi* et al. *(2000)***
54	1prc	2.3		Deisenhofer *et al.* (1985)
55	1ogv	2.4		Katona *et al.* (2003)
56	1aig*	2.6		Chang *et al.* (1991)
57	1pss	3.0		Yeates *et al.* (1987)
58	2rcr	3.1		Chang *et al.* (1991)
			Light harvesting complexes	
59	***1nkz***	***2.0***		***Papiz* et al. *(2003)***
60	1lgh*	2.4		Koepke *et al.* (1996)
61	1kzu	2.5		McDermott *et al.* (1995)
			Fumarate reductase/Succinate dehydrogenase	
62	***1qla/b***	***2.2***		***Lancaster* et al. *(1999)***
63	1l0v*	3.3		Iverson *et al.* (1999)
64	1nek/n	2.9		Yankovskaya *et al.* (2003)
			ATP synthases	
65	1c17*	N/A	NMR, partial structure	Girvin *et al.* (1998)
66	1qo1	3.9	Back-bone only	Stock *et al.* (1999)
			Formate dehydrogenase	
67	***1kqf/g****	***2.8***		***Jormakka* et al. *(2002)***
			Nitrate reductase	
68	***1q16***	***1.9***		***Bertero* et al. *(2003)***
			Adenine nucleotide carrier	
69	***10kc***	***2.2***		***Pebay-Peyroula* et al. *(2003)***
			Cytochrome C oxidase	
70	***1ehk****	***2.4***		***Soulimane* et al. *(2000)***
71	1occ	2.8		Tsukihara *et al.* (1996)
72	1ar1	2.8		Iwata *et al.* (1995)
			Ubiquinol oxidase	
73	***1fft****	***3.5***		***Abramson* et al. *(2000)***

Table 3.1: *continued*

	PDB	Resolution	Notes	Reference
			Cytochrome Bc1	
74	*1bgy**	*2.8-3.0*		***Iwata* et al. *(1998)***
75	1qcr	2.9		Xia *et al.* (1997)
76	1bcc	3.2		Zhang *et al.* (1998)
77	1ezv	2.3	Partial structure	Hunte *et al.* (2000)

Table 3.1: Polytopic α-helical TM protein structures available in January 2004. Structures that were included in the analysis performed here are indicated in red, in comparision with the highest resolution (non-mutant, complete, uncomplexed) homologues available in bold italics. Each horizontal line divides the structures into groups of homologues sharing 25% of greater sequence identity. * indicates the 18 structures added to the dataset in June 2002. All others were added when the dataset was updated in January 2004.

In retrospect it was discovered that several of the proteins in the final dataset were not the very highest resolution structures available (excluding partial and mutant structures and those crystallised in complex with an antibody). This is shown in Table 3.1. It has generally arisen because, when the dataset was updated, as new structures became available, new proteins were excluded if a homologue was already present in the dataset, even if the new protein was of better resolution. However, the average resolution of the proteins analysed was 2.93Å and this would only have been increased to 2.82Å, had all of the highest resolution structures been used. Given the fact that the analyses presented in this chapter are all at the residue, rather than atom, level, and so not dependent on side-chain conformations, this small reduction is resolution is unlikely to have significantly affected the results.

All results were obtained using all 24 proteins listed in Table 3.3, except the comparison of lipid-tail-spanning and head-group-spanning residues (Section 3.3.5.2) and the calculation of position-specific residue distributions (Section 3.3.5.1), which were performed using only 18 of the 24 dataset proteins. This is because this work was performed in June 2002, when only 18 of the 24 proteins were available. These 18 proteins are indicated in Table 3.3.

A viral fusion protein (2siv) is excluded from the calculations, due to its extremely unusual amino acid composition. These unusual properties are likely to be caused by the ability of the protein exist in both aqueous and membrane-spanning environments, unlike

a typical TM protein (Malashkevich *et al.*, 1998).

18 water-soluble proteins for comparison with the membrane proteins were randomly extracted by text searching of the PDB (update version November 2002, containing 19311 structures) with the keyword 'protein'. Homologous proteins were removed using the same criteria as for the TM proteins.

3.2.3 Identification of TM Helices from 3-dimensional co-ordinates by PSlice

A program, named PSlice, was developed in Perl to identify TM helices from 3-dimensional co-ordinate data. This enabled all residues in each structure to be classified as either lipid-tail-spanning, head-group-spanning or non-membrane-spanning. This step is important even in cases where annotation of TM helix positions is available, to allow the published TM helix positions to be verified and ensure that they are consistently defined across all proteins in the dataset.

PSlice has 4 distinct stages. These stages are illustrated in Figure 3.2 and described in more detail in this section.

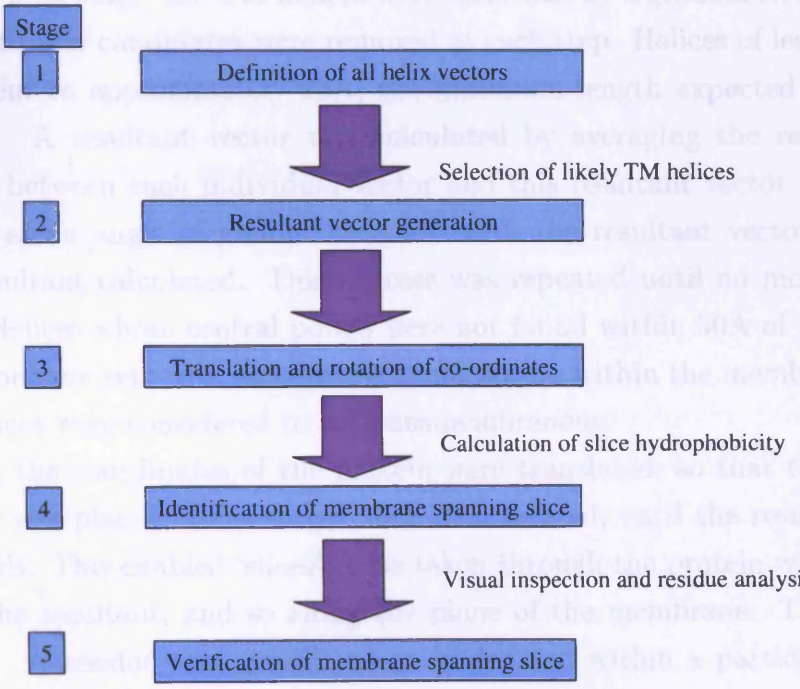


Figure 3.2: Flow diagram showing the stages involved in the program PSlice.

In stage 1, all of the α-helices in the protein were defined and assigned to vectors

by the algorithm ProSEC (Slidel, 1997). While ProSEC was designed to automatically assign the handedness of a motif, its output can be set to give a set of vectors defining the start and end point of each α-helix and β-sheet. ProSEC uses the assignment of residues to secondary structure classes made by DSSP (Kabsch & Sander, 1983) to define the positions of helices and strands. DSSP assigns residues to these secondary structure classes on the basis of the strength of hydrogen bonds that can be inferred from a simple electrostatic model. To split adjacent helices that have been joined, the angles between sequential pairs of residues are determined, and breaks or joins are introduced between the secondary structure elements as appropriate.

The vectors defined by ProSEC were analysed to ensure no curved helices had been split into two or more shorter straight ones. For this, those helices with end points within 3 residues of each other and at an angle of less than 40° to one another were considered to be part of the same helix. The vector through the new helix was determined as a straight line which connected the furthest ends of the two shorter helices, as defined by the position of each residue concerned in the sequence. This was designed to allow for detection of proline-kinked helices, which have been estimated to kink by up to 40° (von Heijne, 1991; Woolfson & Williams, 1990; Nilsson *et al.*, 1998; Chang *et al.*, 1999).

During the next stage, the TM helices were identified by a gradual iterative process, in which the least likely candidates were removed at each step. Helices of less than 15 amino acids, equivalent to approximately 23Å, the minimum length expected for TM helices, were removed. A resultant vector was calculated by averaging the remaining helices, and the angle between each individual vector and this resultant vector was determined. Those helices at an angle of greater than 50° with the resultant vector were removed, and a new resultant calculated. This process was repeated until no more vectors could be removed. Helices whose central points were not found within 50Å of the centre of the resultant vector were removed, since these could not be within the membrane plane. The remaining helices were considered to be transmembranous.

In stage 3, the coordinates of the protein were translated, so that the optimised resultant vector was placed at the origin, and then rotated, until the resultant vector ran along the y-axis. This enabled 'slices' to be taken through the protein with ease, perpendicularly to the resultant, and so along the plane of the membrane. This is illustrated in Figure 3.3. A residue was considered to be located within a particular slice only if its Cα-carbon atom was found within the correct co-ordinate range. Slices of various thicknesses were used in order to identify the one that most corresponds to the thickness of the lipid-tail-spanning region of a membrane. It was then possible to select the slice through which the membrane was most likely to lie, due to the maximal hydrophobicity of its surface accessible residues. (Surface accessible residues were defined as described in

Section 3.2.4.2).

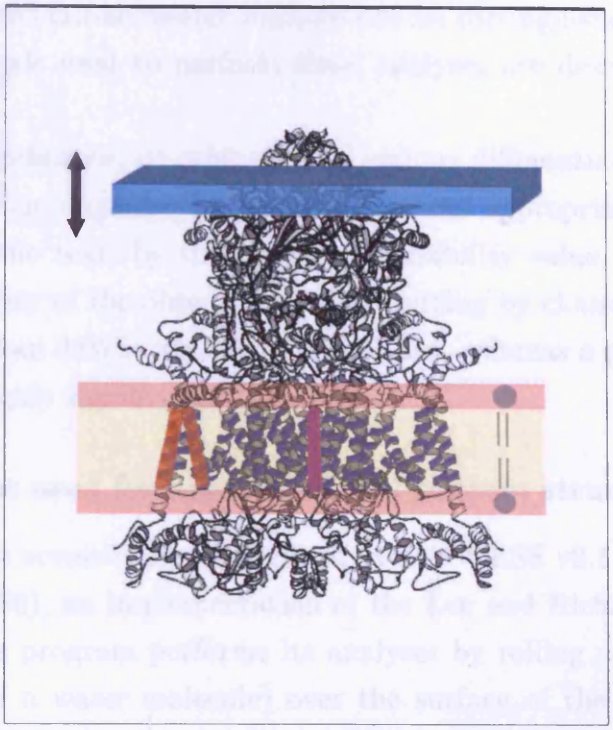


Figure 3.3: Schematic diagram illustrating the method used by PSlice to identify the membrane-spanning slice for cytochrome Bc1 (PDB code 1bgy). Shows the predicted position of the membrane (red and yellow faint overlay), two transmembrane helices (red bars) and the resultant vector (purple bar). The turquoise box illustrates the slicing method used by the program PSlice to calculate the surface hydrophobicity of slices perpendicular to the resultant vector.

Hydrophobicity was calculated, in the final stage of the algorithm, and throughout this work, using the White and Wimley (WW) scale (Wimley *et al.*, 1996; Jayasinghe *et al.*, 2001), which is described in Section 3.2.4.3. Lastly, the location of the membrane-spanning region, as defined by PSlice, was validated by visual inspection (Results, Section 3.3.2) and by analysis of the distribution of particular residue types (Results, Section 3.3.5.1).

3.2.4 Computational analysis of TM protein structure

3.2.4.1 Overview

Both sequence and geometric characteristics of the dataset TM helices were studied. Geometric data includes both helix lengths and packing angles, which were determined by

PSlice. Sequence-based data consist of residue distributions, hydrophobicity and conservation analysis. These were analysed in order to determine by which sequence-derived criteria buried and lipid-tail-accessible residues can be distinguished. The programs and the hydrophobicity scale used to perform these analyses are described in the following sections.

The statistical significance, or otherwise, of various differentials between groups was determined by paired or unpaired Student's T-tests as appropriate. Significance is indicated, throughout the text, by the use of a probability value, P. A certain P value indicates the probability of the observed result occurring by chance alone. In this work, a probability of less than 0.05 is considered significant, whereas a probability of less than 0.001 is considered highly significant.

3.2.4.2 Algorithms used for analysis of TM protein structure

NACCESS Residue accessibility was defined by NACCESS v2.1.1 (©, S. Hubbard and J. Thornton, 1992-1996), an implementation of the Lee and Richards algorithm (Lee & Richards, 1971). This program performs its analyses by rolling a sphere of radius 1.4Å (equivalent to that of a water molecule) over the surface of the protein structure and determining to what degree residues are accessible to it. For this work, residues with a relative accessibility score of greater than 5% are considered to be lipid-tail-accessible, whereas those with a score of less than 5% are considered to be buried. This value had previously been used successfully by others (Valdar & Thornton, 2001; Bartlett *et al.*, 2002).

PSI-BLAST PSI-BLAST (Altschul *et al.*, 1997) is described in detail in Section 1.3.1 in Chapter 1. For this work, PSI-BLAST was run, for a maximum of 20 iterations or until convergence, using a threshold of $1e^{-40}$. A relatively high threshold was used to select only close homologues, for which function is likely to be conserved.

SCORECONS The calculation of residue conservation from PSI-BLAST-derived alignments was performed by SCORECONS (Valdar & Thornton, 2001), as described in Section 1.3.2 in Chapter 1.

3.2.4.3 White and Wimley hydrophobicity scale

Throughout this chapter the White and Wimley (WW) scale (Wimley *et al.*, 1996; Jayasinghe *et al.*, 2001) was used to assess the hydrophobicity of residues. This scale takes into account the solvation energy of both the amino acid side chain and the peptide backbone.

The WW scale is derived from the partitioning of peptides, containing the residue under study, into n-octanol. The hydrophobicity score of each residue according to the WW scale, calculated from the partitioning coefficient, is shown in Figure 3.4. This scale was selected, firstly, since it has been shown to predict TM helices with an accuracy of greater than 99% from sequence (Jayasinghe *et al.*, 2001). (While the criteria used to define success were relatively relaxed (an overlap of three or more residues between a predicted and known TM helix), the method did perform better than several comparable scales, the GES (Engelman *et al.*, 1986), KD (Kyte & Doolittle, 1982) and EC (Eisenberg *et al.*, 1982a) scales, using the same criteria). Secondly, initial studies suggested that the WW scale was also effective for the current work, since a distinct peak of hydrophobicity in the residues along the protein surface could be identified.

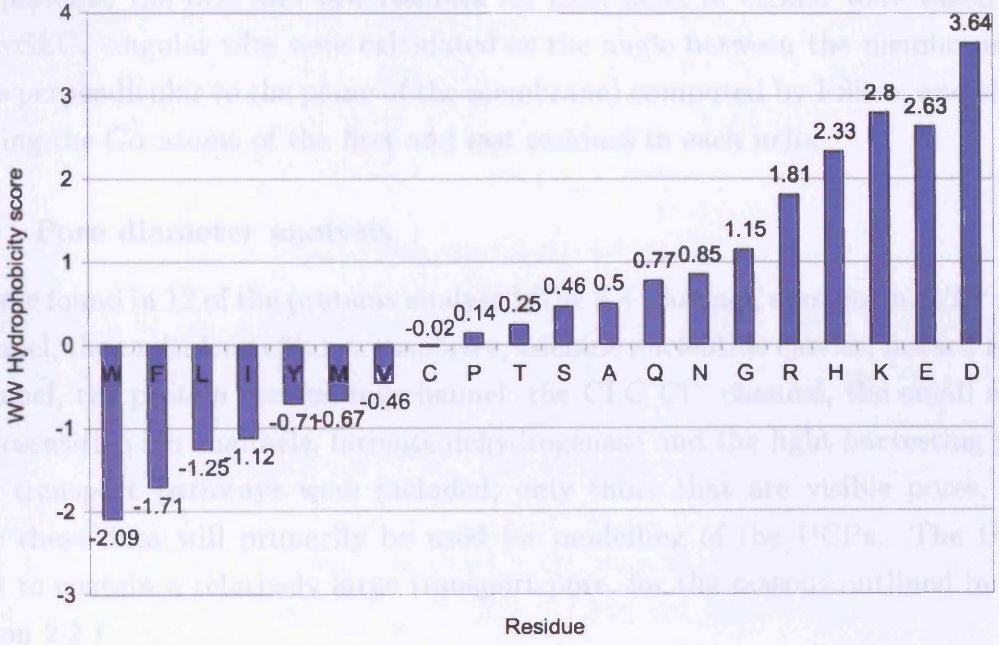


Figure 3.4: Graph indicating the hydrophobicity score of each residue, according to the WW scale (Wimley *et al.*, 1996; Jayasinghe *et al.*, 2001).

3.2.4.4 Assignment of residues to classes for analysis

Lipid-tail-spanning residues were defined as those with their Cα atom within the 30Å lipid-tail-spanning slice identified by PSlice, according to the 3-dimensional coordinates. Head-group-spanning residues were defined as those residues with their Cα atom within either of the 15Å head-group-spanning slices flanking the lipid-tail-spanning slice, as illustrated in Figure 3.1. The residues within the lipid-tail-spanning and head-group-spanning regions

were defined as membrane-spanning. All other residues were classified as non-membrane-spanning and were excluded from the analyses. The membrane-spanning residues were further subdivided into lipid-tail-accessible, lipid-tail-spanning buried, head-group-accessible and head-group-spanning buried groups, as determined by their accessible surface area calculated by NACCESS.

3.2.4.5 Comparison of the secondary structure characteristics of TM and water-soluble proteins

For each TM helix defined by ProSEC, the helix termini were defined as the two residues furthest apart in the helix sequence, for which the Cα atom was found within the TM lipid-tail-spanning slice defined by PSlice. For water-soluble proteins, and non-TM helices in TM proteins, the first and last residues for each helix or strand were taken directly from ProSEC. Angular tilts were calculated as the angle between the membrane normal (the line perpendicular to the plane of the membrane) computed by PSlice, and the vector connecting the Cα atoms of the first and last residues in each helix.

3.2.4.6 Pore diameter analysis

Pores were found in 12 of the proteins analysed (the K+ channel, aquaporin, ATP synthase H^+ channel, the multidrug efflux transporter, adenine nucleotide carrier, inward rectifying K^+ channel, the protein conducting channel, the CLC Cl^- channel, the small and large mechanosensitive ion channels, formate dehydrogenase and the light harvesting protein). Not all transport pathways were included, only those that are visible pores. This is because these data will primarily be used for modelling of the UCPs. The UCPs are thought to contain a relatively large transport pore, for the reasons outlined in Chapter 2, Section 2.2.1.

The pore diameter is taken to be the distance, according to the 3-dimensional coordinates, between the Cα atoms of the two pore-lining residues at any particular height that are furthest apart. Hence the pore diameters are somewhat over-estimated since they do not take into account the volume of the side chains. The pore diameters used in this work are the average of 3 diameters, calculated at different heights in the pore. Both functional and non-functional pores (defined in Section 3.2.4.12) are included in the analysis of pore diameter. The pore diameters of the UCP models from Chapter 2 are estimated based on the idea that the diameter of a TM helix is 10Å.

3.2.4.7 Distribution of residue types across the membrane-spanning regions

The propensities calculated in this work are both normalised with respect to the total number of residues within each slice, and with respect to the total number of each residue type. The propensity of residue X in slice S is therefore found using Formulae 3.1-3.3:

$$\%\text{ of X in slice S} = \frac{\text{number of X in slice S}}{\text{number of all residues in slice S}} \tag{3.1}$$

$$\text{Average \% of X in all slices} = \frac{\text{total number of X in TM region of all proteins}}{\text{total number of all residues in TM region of all proteins}} \tag{3.2}$$

$$\text{Propensity of X in slice S} = \frac{\%\text{ of X in slice S}}{\text{Average \% of X in all slices}} \tag{3.3}$$

As a result, a propensity of 1 represents the average proportion of residues of a certain type within the whole TM region. A propensity of greater than 1 indicates a greater than average proportion of that residue, and hence an enrichment at that location. Conversely, a propensity of less than 1 indicates a lower than average proportion, and hence that the residue is disfavoured at that location.

3.2.4.8 Comparison of the hydrophobicity of accessible and buried residues

For each TM helix the average hydrophobicity score on the White and Wimley scale (Wimley *et al.*, 1996; Jayasinghe *et al.*, 2001) of the lipid-tail-accessible residues was compared to that for the buried residues. Lipid-tail-accessible and buried residues were defined as described in Section 3.2.4.2. The hydrophobicity scale used is described in Section 3.2.4.3.

The calculation was performed separately for both the lipid-tail-spanning and head-group-spanning regions of the helices. In order to prevent bias of the results towards the characteristics of certain chains, if a protein contained identical chains, only the helices belonging to one of these were analysed. The chains that were excluded from analysis are shown in Table 3.2. This was necessary because the structures of chains with identical sequences are usually identical. Chains with identifiers (taken from the PDB file) nearest to the start of the alphabet were arbitrarily retained at the expense of those nearer the end (i.e. chain A was included at the expense of chain B, in a homodimeric protein consisting of the two identical chains A and B).

For the lipid-tail-spanning region, of the 455 TM helices in the dataset of 24 TM proteins, 217 remained after identical chains were removed. 14 helices contained either no accessible or no buried residues, leaving 203 for analysis. Similarly, for the head-group-spanning region, of the 429 non-identical TM helices in the dataset, 224 helices contained either no accessible or no buried residues, leaving 215 for analysis.

Protein	Included chains	Excluded chains
1ehk	ABC	
1c17	AM	BCDEFGHIJKL
1jb0	ABCDEFGHIJKLMX	
1iwg	A	
1lgh	A	
1eul	A	
1aig	HLM	NOP
1l0v	ABCD	MNOP
1bgy	ABCDEFGHIJK	MNOPQRSTUVW
1bl8	A	BCD
1um3	ABCDEFGH	NMOPQRST
1l7v	AB	CD
1l9h	A	B
1fft	ABC	FGH
1h6i	A	BCD
1mxm	A	BCDEFG
1msl	A	BCDE
1okc	A	
1p7b	A	B
1pv6	A	B
1q16	ABC	
1rhz	ABC	
1kpl	A	BCD
1kqf	ABC	DEFGHI

Table 3.2: Identical chains included and excluded from the analyses of helix hydrophobicity and conservation.

3.2.4.9 Comparison of the sequence conservation of accessible and buried residues

PSI-BLAST (Section 3.2.4.2) was used to identify sequence homologues for each of the proteins of known structure. The prediction of residue conservation amongst these homologues was performed by SCORECONS (Section 3.2.4.2).

For each TM helix, the average conservation score of the lipid-tail-accessible residues was compared to that for the buried residues. The calculation was performed separately for both the lipid-tail-spanning and head-group-spanning regions of the helices. As for the hydrophobicity analysis in Section 3.2.4.8, helices from identical chains were removed (Table 3.2). For the lipid-tail-spanning region, of the 455 TM helices in the dataset,

217 remained after identical chains were removed. (In the case of proteins containing identical chains, the chains with identifiers closest the the beginning of the alphabet were arbitrarily kept at the expense of homologues with later identifiers). 8 of these helices contained either no accessible or no buried residues, and insufficient homologues were found to derive conservation scores for another 49 helices, leaving 160 for analysis. Similarly, for the head-group-spanning region, of the 439 non-identical TM helices in the dataset, 162 helices contained either no accessible or no buried residues, and insufficient homologues were found for 112 helices, leaving 165 for analysis.

3.2.4.10 Comparison of the preferences of particular residues for lipid-tail-accessible or buried positions

The preference that a residue, X, shows for lipid-tail-accessible positions can be calculated using the following formula:

$$\text{Propensity of X in environment E} = \frac{\%\text{ of X in lipid-tail-accessible positions}}{\text{Average \% of all residues in lipid-tail-accessible positions}} \quad (3.4)$$

3.2.4.11 Hydrogen bond analysis

Hydrogen bonding partners were analysed for the 1047 observed lipid-tail-accessible charged residues and the 1202 lipid-tail-accessible polar residues in the dataset. Charged residues are arginine, lysine, glutamate, histidine and aspartate. Polar residues are serine, threonine, asparagine, glutamine, cysteine. The remaining residues are classified as hydrophobic. Hydrogen bonds were detected by HBPlus v3.0 (McDonald & Thornton, 1994) and classified using a Perl script developed for the purpose. HBPlus identifies hydrogen bonds using the inter-atom distance and angle criteria defined by Baker & Hubbard (1984) and confirmed in an analysis of bond geometries (Thornton *et al.*, 1993). Main-chain/main-chain hydrogen bonds were excluded from the analysis, in order to analyse only the hydrogen bonding patterns of the residue side-chains. Intrahelical hydrogen bonds are those that are formed between 2 residues found 3 or 4 positions apart on the same chain. All other hydrogen bonds are classed as interhelical. Only hydrogen bonds to protein are detected by HBPlus so bonds to head-groups are inferred from visual inspection of the structures.

Snorkelling residues are residues with Cα atoms within the lipid-tail-spanning region that adopt a conformation to permit their positively charged groups to reside in the head-group-spanning region (described in Section 3.3.8). Positively charged residues are classified as snorkelling if their Cα atoms are less than 8Å from the head-group-spanning region. (The length of the lysine side-chain was estimated at approximately 8Å, from bond length calculations).

3.2.4.12 Analysis of pore-lining residues

The characteristics of pore-lining residues within the lipid-tail-spanning or head-group-spanning regions were compared to those of buried and lipid-tail-accessible residues. Not all residues involved in transport were included, only those lining visible pores. This is because the UCPs are thought to contain a relatively large transport pore, for the reasons outlined in Chapter 2, Section 2.2.1. Only functional pores (defined as those through which transport is known to occur) were included in the analysis since a preliminary study showed differences between the residue composition of functional and non-functional pores. (Non-functional pores were more hydrophobic, probably because they are packed with phospholipids in the native structure). A functional pore was present within each protomer of the K+ channel, aquaporin, the multidrug efflux transporter, the adenine nucleotide carrier, the inward rectifying K+ channel, the protein conducting channel, the Clc chloride channel and the small and large mechanosensitive ion channels. The channels found in formate dehydrogenase, ATP synthase and the light harvesting protein, as well as the inter-protomer pores in aquaporin and the multidrug efflux transporter, were classed as non-functional and excluded from the analysis of pore-lining residue types. Functional pores are indicated in Figures 3.12 and 3.13.

3.3 Results

3.3.1 The dataset of transmembrane protein structures

At the time of dataset assembly (January 2004) there are 24 non-homologous α-helical polytopic TM proteins with known 3-dimensional structures, and these are listed in Table 3.3. The proteins average approximately 125kDa in size, although some are small single polypeptide chain proteins, whereas others are very large protein complexes consisting of up to 20 chains. They contain 455 TM helices in total, with an average of 19 TM helices per protein. The smallest protein, the adenine nucleotide carrier, has 6 TM helices, whereas the largest, the multidrug efflux transporter, has 36.

It is accepted that these proteins are unlikely to represent a random sample of the total proteome, due to greater scientific interest or ease of expression of some classes of proteins, such as prokaryotic proteins and the mitochondrial electron transport proteins, leading to their over-representation in the structural databases. This may lead to the dataset becoming biased by characteristics specific for these types of protein. However, since the dataset is also likely to be biased towards scientifically and medically important proteins, the implications of this bias may be reduced. In addition, the number of available TM protein structures is too small to allow rejection of a protein from the analysis based on

anything other than close sequence similarity.

Name	PDB	Reference	Origin	No. Hom	O/S	No. TMH
Fumarate reductase (Complex II)	1l0v	Iverson *et al.* (1999)	Bacterial	1	1	12
Cytochrome Bc1 Complex III)	1bgy	Iwata *et al.* (1998)	Eukaryotic	3	2	28
Cytochrome C oxidase (Complex IV)	1ehk	Soulimane *et al.* (2000)	Bacterial	2	1	15
F1F0 ATP synthase Chain C (Complex V)	1c17	Girvin *et al.* (1998)	Bacterial	1	1	28
Photosystem I	1jb0	Jordan *et al.* (2001)	Bacterial	3	1	28
Formate dehydrogenase	1kqf	Jormakka *et al.* (2002)	Bacterial	0	3	5
Multidrug efflux transporter	1iwg	Murakami *et al.* (2002)	Bacterial	0	3	12
Light harvesting protein	1lgh	Conroy *et al.* (2000)	Bacterial	1	1	16
Photosynthetic reaction centre	1aig	Chang *et al.* (1991)	Bacterial	2	1	11*
Calcium ATPase	1eul	Toyoshima *et al.* (2000)	Eukaryotic	0	1	10
CLC Cl^- channel	1kpl	Dutzler *et al.* (2002)	Bacterial	2	2	13

Table 3.3: *continued*

Name	PDB	Reference	Origin	No. Hom	O/S	No. TMH
K^+ channel	1bl8	Doyle *et al.* (1998)	Bacterial	0	1	8
ABC Transporter	1l7v	Locher *et al.* (2002)	Bacterial	2	2	9
Rhodopsin	1l9h	Palczewski *et al.* (2000)	Eukaryotic	3	2	7
Ubiquinol oxidase	1fft	Abramson *et al.* (2000)	Bacterial	0	1	50
Aquaporin	1h6i	de Groot *et al.* (2001)	Eukaryotic	1	1	32
Small mechanosensitive channel (MscS)	1mxm	Bass *et al.* (2002)	Bacterial	0	7	3
MscL mechanosensitive channel (MscL)	1msl	Chang *et al.* (1998)	Bacterial	0	1	10
Adenine nucleotide carrier	1okc	Pebay-Peyroula *et al.* (2003)	Eukaryotic	0	1	6
Inward rectifier K^+ channel	1p7b	Kuo *et al.* (2003)	Bacterial	0	4	8
Lactose permease	1pv6	Abramson *et al.* (2003)	Bacterial	0	1	12

Table 3.3: *continued*

Name	PDB	Reference	Origin	No. Hom	O/S	No. TMH
Nitrate reductase	1q16	Bertero *et al.* (2003)	Bacterial	0	2	10
SecYEβ protein conducting channel	1rhz	Van den Berg *et al.* (2004)	Bacterial	0	1	11
Cytochrome B6F	1um3	Kurisu *et al.* (2003)	Bacterial	0	2	24

Table 3.3: Non-homologous polytopic α-helical membrane proteins with known 3-dimensional structure. Also given are PDB code (PDB), origin, number of homologues of known 3-dimensional structure (No. Hom), oligomeric state (O/S) and number of TM helices in the native protomer (No. TMH). The number of TM helices per protomer is given according to either SwissProt (Bairoch & Apweiler, 1997) annotation or visual inspection of protein structures if this is unavailable. The oligomeric state of the native protein is taken from the PQS server (K. Henrick, *http://pqs.ebi.ac.uk/*). *The photosynthetic reaction centre contains a chain that spans the membrane with a single TM helix. In retrospect this chain should have been removed from the dataset since it is not polytopic. Cofactors were not removed from the structures.

3.3.2 Location of TM helices from 3-dimensional coordinates

A clear trough in hydrophobicity was visible when surface-accessible residues were analysed in slices taken along the plane of the membrane calculated by PSlice. This is illustrated in Figure 3.5 for slice thicknesses ranging from 20Å to 50Å, for cytochrome Bc1 (PDB code 1bgy). A 30Å thick slice was selected since it gave the deepest, sharpest trough in hydrophobicity score in an arbitrary sample of 10 of the dataset proteins, allowing the TM slice to be identified as accurately as possible. This was in agreement with the majority of thicknesses given for the membrane lipid-tail environment in the literature. Whilst throughout this work a membrane lipid-tail thickness of 30Å was used, this figure is thought to vary by up to 7.5Å, (Bretscher & Munro, 1993; Killian, 1998) due to differing lipid composition and cholesterol content. This may have lead to the inclusion of some head-group-spanning residues in the lipid-tail-spanning region, or vice versa, particularly for some proteins such as the ATP synthase.

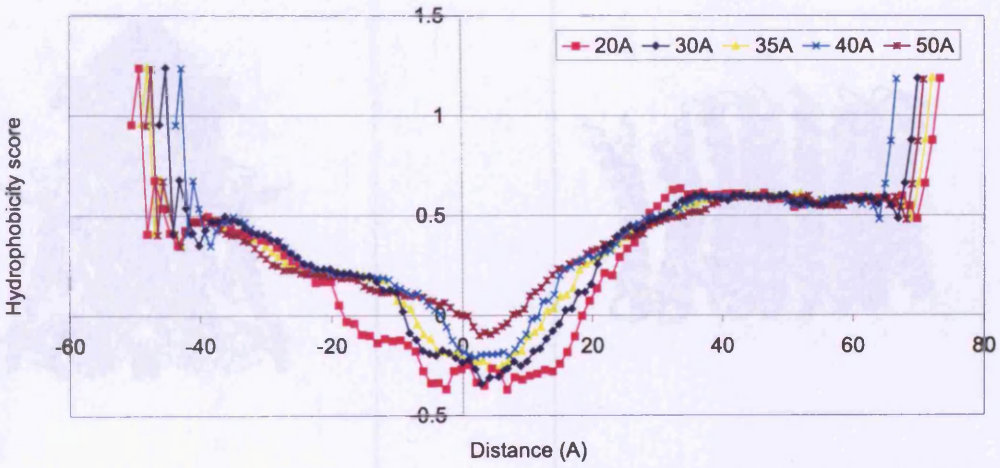


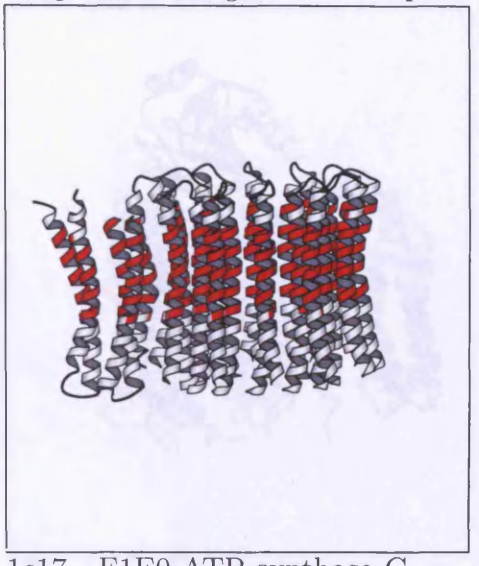
Figure 3.5: Distribution of hydrophobicity of all surface residues of cytochrome Bc1 (PDB code 1bgy), using slices of varying thickness along the plane of the membrane. The lower the hydrophobicity score, the greater the hydrophobicity. The hydrophobicity score for each slice was plotted on the x-axis at the centre of that slice.

The position of the TM slice was verified by visual inspection of the structure, as shown in Figures 3.6–3.8, and by the distributions of residues described in Section 3.3.5.1. Another valuable method to verify the predicted position of the TM slice would have been to combine the results from homologous proteins, since these are likely to have a very similar structure.

1aig - Photosynthetic reaction centre

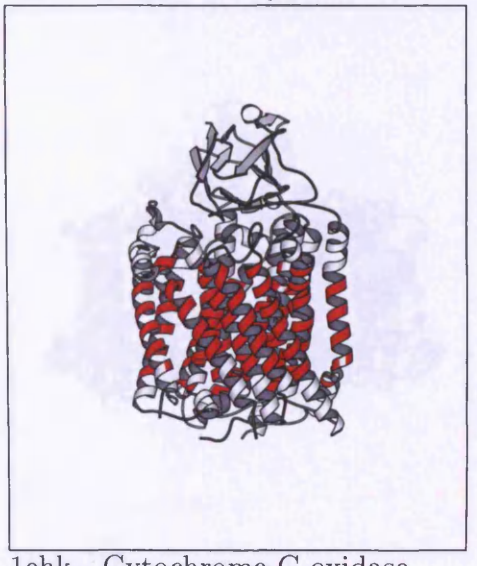
1iwg - Multidrug efflux transporter

1bgy - Cytochrome Bc1



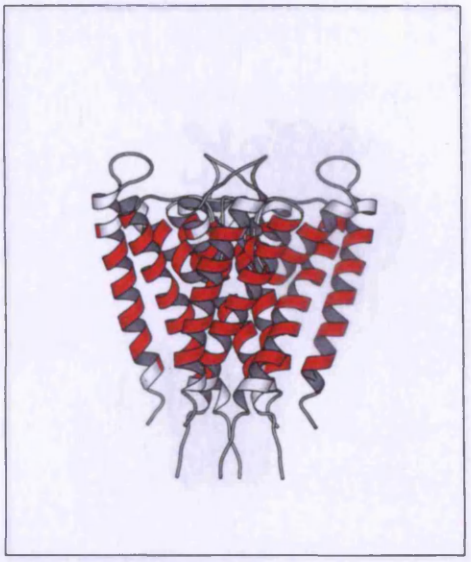
1c17 - F1F0 ATP synthase C

1kqf - Formate dehydrogenase

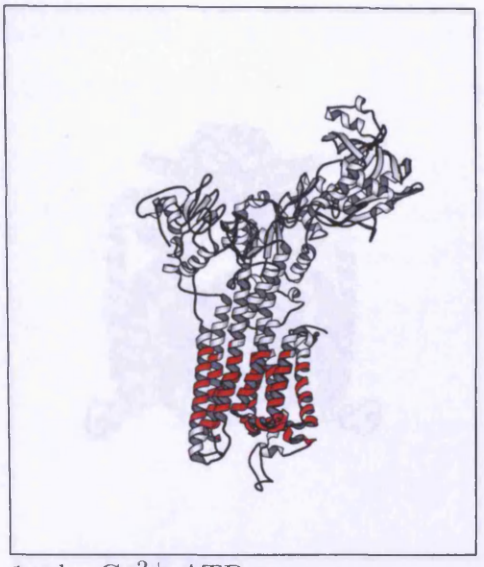


1ehk - Cytochrome C oxidase

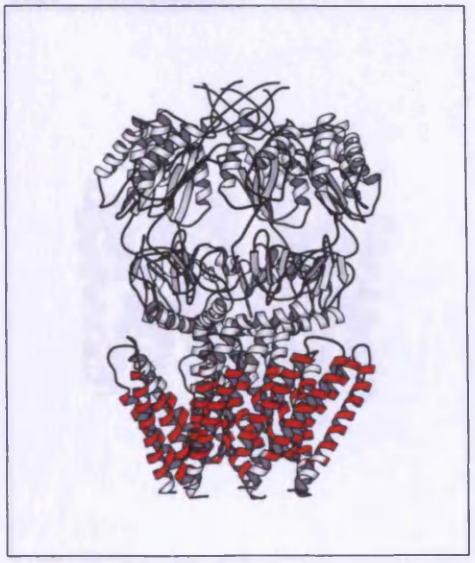
Figure 3.6: Structures of the dataset proteins showing the TM slice identified by PSlice in red.



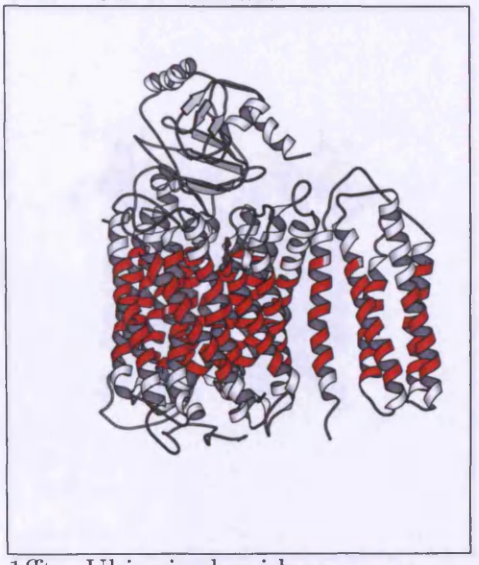
1bl8 - K^+ channel



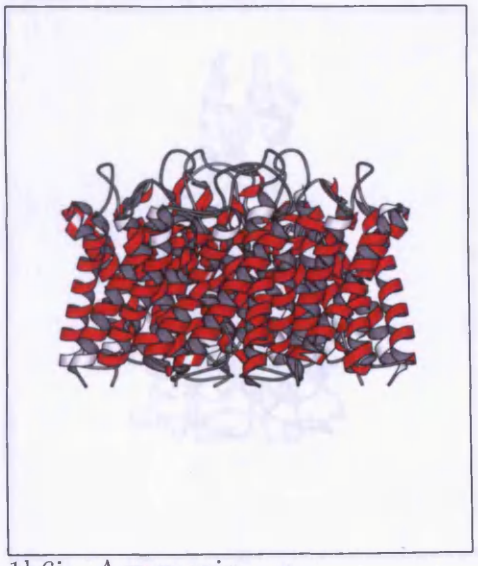
1eul - Ca^{2+} ATPase



1mxm - MscS channel



1fft - Ubiquinol oxidase



1h6i - Aquaporin

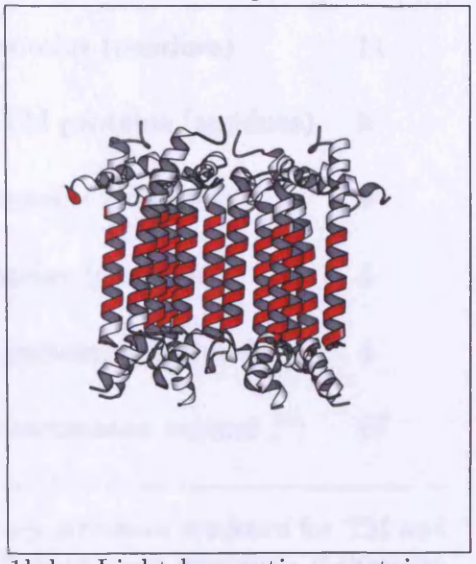
1jb0 - Photosystem I

Figure 3.7: Structures of the dataset proteins showing the TM slice identified by PSlice in red. This and all similar figures were produced using Pymol (©DeLano Scientific, 2004).

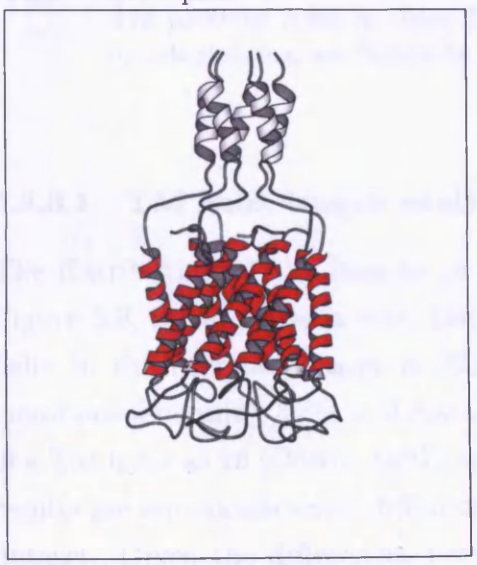
1l0v - Fumarate reductase

1l7v - ABC Transporter

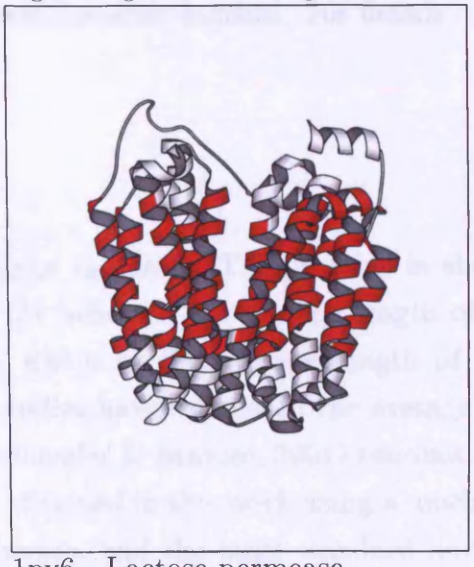
1l9h - Rhodopsin



1lgh - Light harvesting protein



1msl - MscL channel



1pv6 - Lactose permease

Figure 3.8: Structures of the dataset proteins showing the TM slice identified by PSlice in red.

3.3.3 A comparison of the secondary structure characteristics of membrane and water-soluble proteins

Various statistics concerning the number, length and angle of α-helices and β-sheets in TM proteins are given in Table 3.4, in comparison with the corresponding values for water-soluble proteins. These data are derived as described in the methods section. The results are not significantly different from those obtained previously with smaller datasets (Bowie, 1997; Ulmschneider & Sansom, 2001).

Average number of TM helices	19
Average TM helix length (residues)	23
Average length of all helices in TM proteins (residues)	11
Average length of non-TM helices in TM proteins (residues)	9
Average length of helices in soluble proteins (residues)	9
Average length of β-sheets in TM proteins (residues)	4
Average length of β-sheets in soluble proteins (residues)	4
Average angular tilt of TM helices to membrane normal (°)	17

Table 3.4: Comparison of various secondary structure statistics for TM and water-soluble proteins. Only α-helical TM proteins are analysed: β-sheets in TM proteins refer to those found in the water-soluble domains. For details of calculations, see Methods Section.

3.3.3.1 TM helix length analysis

The distribution of the lengths of non-TM helices in the 24 TM proteins is shown in Figure 3.9, in comparison with the lengths of TM helices. The average length of a TM helix in the current dataset is 23±6 residues, where as the average length of a non-membrane-spanning helix is 9 residues. Other studies have calculated the average length of a TM helix as 26 (Bowie, 1997) and 27 (Ulmschneider & Sansom, 2001) residues. These results are not significantly different from those obtained in this work using a much larger dataset. Given the differences between the datasets, and the large standard deviations

associated with so few proteins, variation of this order of magnitude between the results of the different studies would be expected.

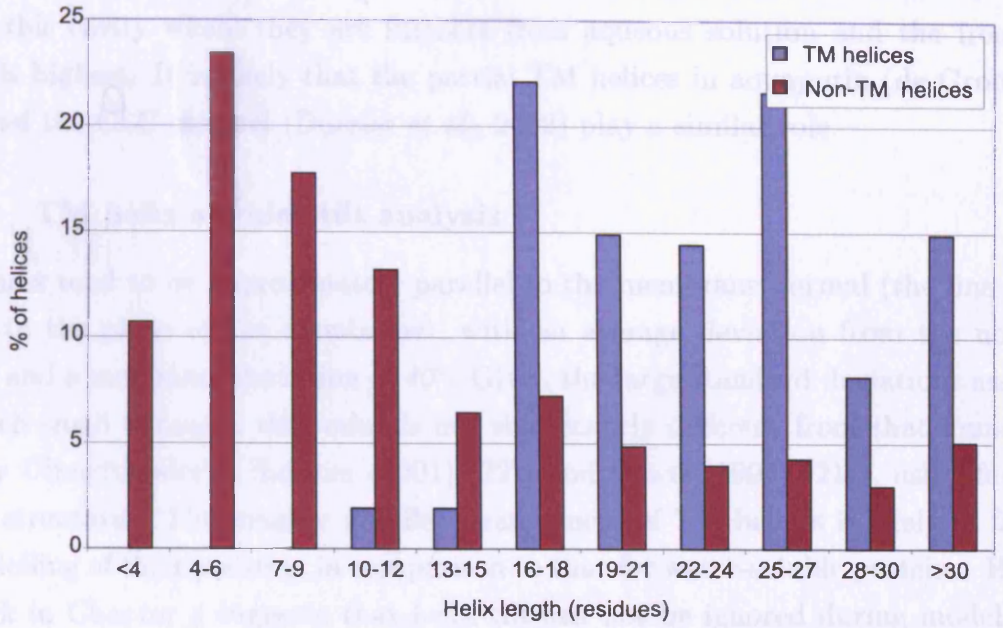


Figure 3.9: Distributions of the lengths of TM helices and non-TM helices in 24 membrane proteins.

As can be seen in Figure 3.9, the majority of helices of greater than 16 residues in TM proteins are TM helices. The remaining helices, and the β-sheets in TM proteins, show a very similar distribution of lengths to the helices and sheets in water-soluble proteins. It can therefore be concluded that, apart from the addition of several longer helices that span the membrane, TM proteins do not appear to differ in their secondary structure composition from non-TM proteins.

3.3.3.2 Partial membrane-spanning TM helices

Even considering that different lipid compositions in eukaryotic and bacterial membranes lead to differing membrane thicknesses (see Section 3.1.1), a helix of 23 residues is likely to be more than adequate to span the membrane bilayer, indicating that most TM helices extend out of the membrane on at least one face. However, several proteins contain helices that do not fully span the membrane, but instead seem to cross only approximately halfway (the K^+ channel (1bl8), aquaporin (1h6i) and the ClC Cl^- channel (1kpl)). In the K^+ channel these partial helices are thought to play vital structural and functional roles by acting as the selectivity filter, reducing the energy barrier for ions crossing and helping to form the characteristic shape of the channel (Doyle *et al.*, 1998). The partial

TM helices point from the extracellular face of the bilayer to a water-filled cavity in the centre of the channel, halfway across the membrane. The dipole created by the partial negative charges on the C termini of these helices is thought to stabilise the translocating ions in this cavity where they are furthest from aqueous solution and the free energy barrier is highest. It is likely that the partial TM helices in aquaporin (de Groot *et al.*, 2001) and the CLC channel (Dutzler *et al.*, 2002) play a similar role.

3.3.3.3 TM helix angular tilt analysis

TM helices tend to be approximately parallel to the membrane normal (the line perpendicular to the plane of the membrane), with an average deviation from the normal of 17±11° and a maximum deviation of 40°. Given the large standard deviations associated with such small datasets, this value is not significantly different from that found previously by Ulmschneider & Sansom (2001) (22°) and Bowie (1997) (21°), using fewer TM protein structures. The roughly parallel arrangement of TM helices is likely to facilitate the modelling of their packing, in comparison to that for water-soluble proteins. However, the work in Chapter 4 suggests that helix tilt can not be ignored during modelling. In contrast, the average deviation of non-TM helices in TM proteins from the membrane normal is 44°. This is illustrated in Figure 3.10. The wide spread of angles for non-membrane-spanning helices indicates that these helices show no tendency to be aligned with the membrane-spanning ones.

It was hoped that information about the length of a TM helix could be used to estimate its angular tilt, and that this may be of use during modelling. The approximate length of the lipid-tail-spanning region of a TM helix can be estimated by maximising the difference in hydrophobicity between the predicted lipid-tail-spanning and head-group-spanning residues. However, as shown in Figure 3.11, there appears to be no significant correlation between the length of a hydrophobic segment and its angle to the membrane normal ($R^2 = 0.08$). This lack of correlation is not due to the fact that the membrane lipid-tail-spanning region often contains TM helices which do not fully span the membrane, since even if these partial-spanning helices are removed, no correlation is observed. However, that partial-spanning helices exist suggests that there is no requirement for the length of a hydrophobic segment to match the thickness of the bilayer. As a result, information about the length of the lipid-tail-spanning part of a helix is unlikely to be of use in the prediction of the angular tilt of the helix in a 3-dimensional model.

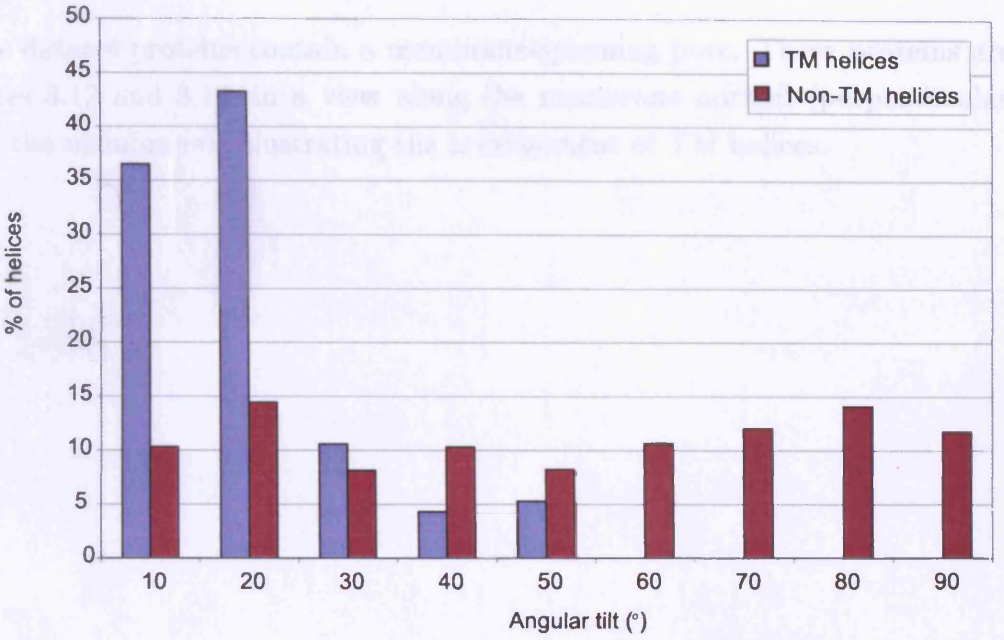


Figure 3.10: Comparison of the distribution of the angular tilt of TM and non-TM helices in 24 membrane proteins from the membrane normal predicted by PSlice.

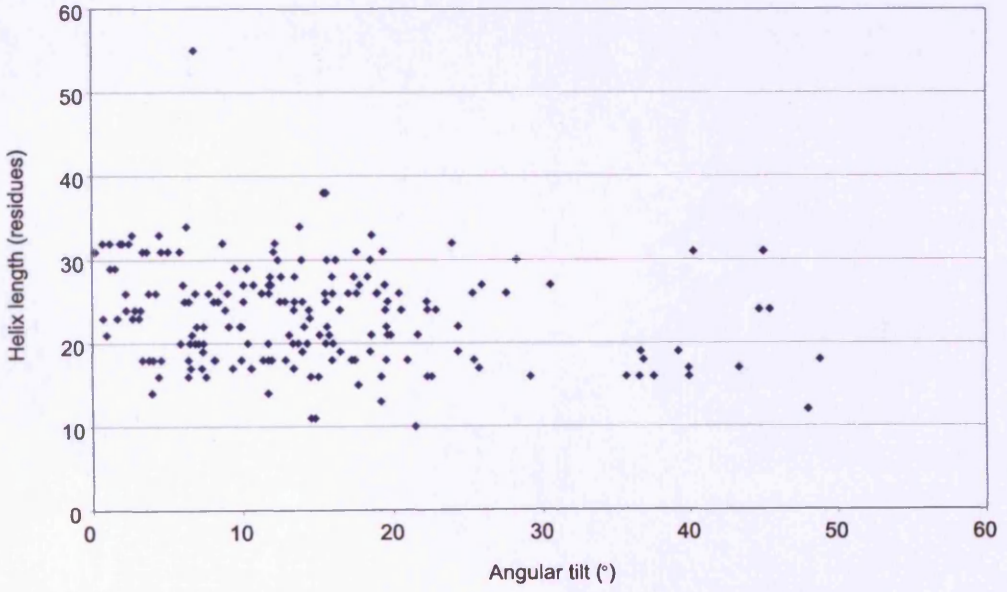
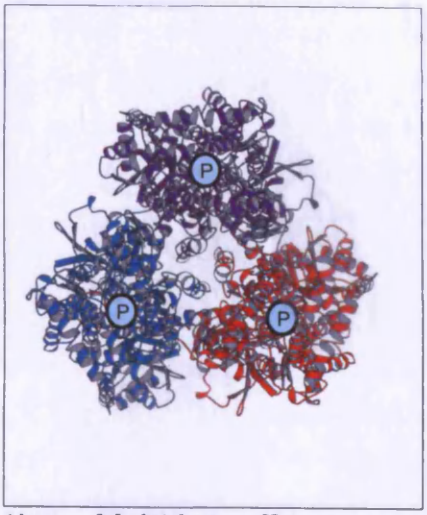


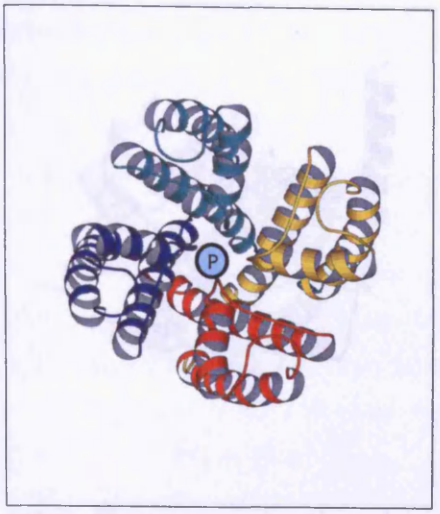
Figure 3.11: Correlation between the lengths and angles from the membrane normal of TM helices in 24 membrane proteins.

3.3.4 Analysis of pore diameter

12 of the dataset proteins contain a membrane-spanning pore. These proteins are shown in Figures 3.12 and 3.13, in a view along the membrane normal (perpendicular to the plane of the membrane), illustrating the arrangement of TM helices.

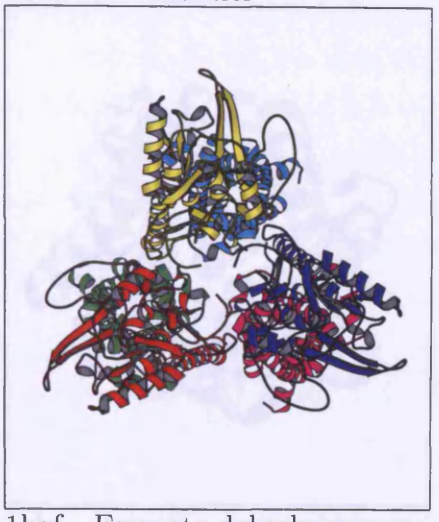


1iwg - Multidrug efflux transporter

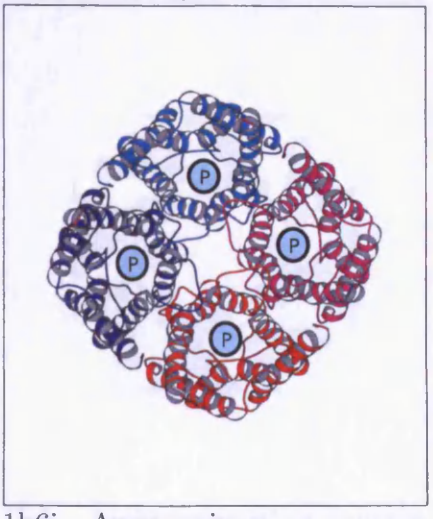


1bl8 - K^+ channel

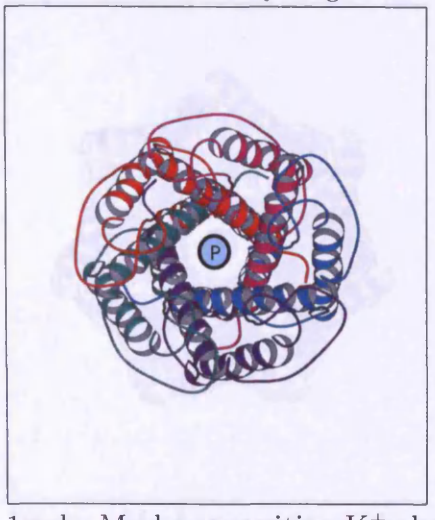
1c17 - ATP synthase subunit C



1kqf - Formate dehydrogenase

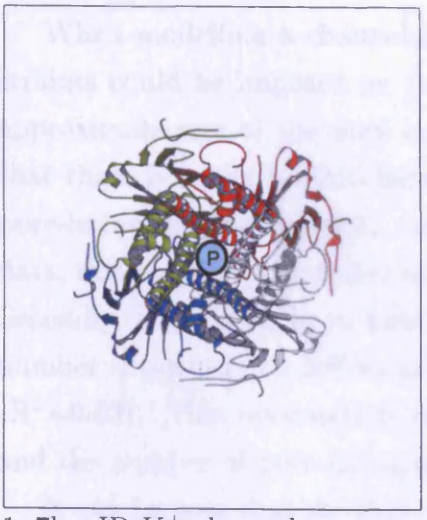


1h6i - Aquaporin

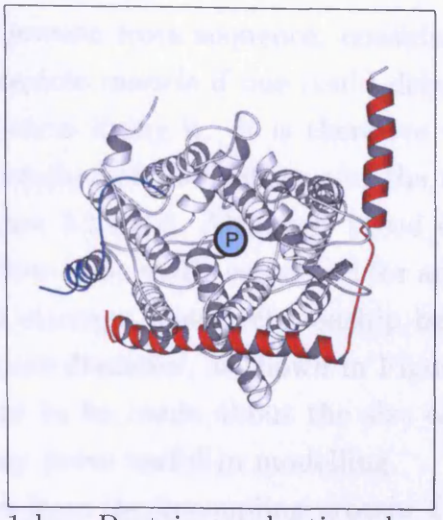


1msl - Mechanosensitive K^+ channel

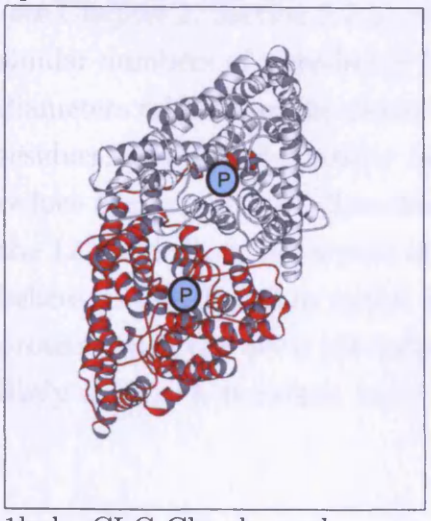
Figure 3.12: Views of the pore-containing TM proteins along the membrane normal, showing the pore and the arrangement of TM helices. This diagram was produced using Molscript ©1997-1998, Per Kraulis. Functional pores are indicated by a pale blue circle containing a 'P'.



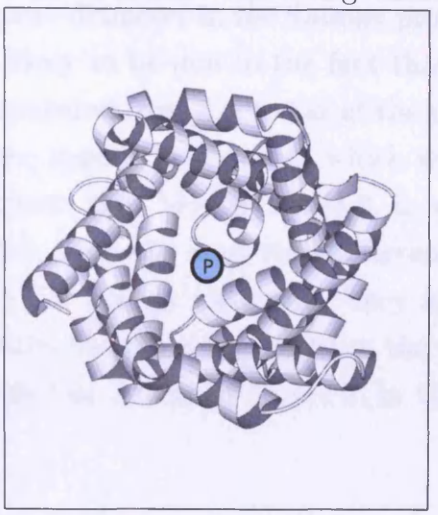
1p7b - IR K^+ channel



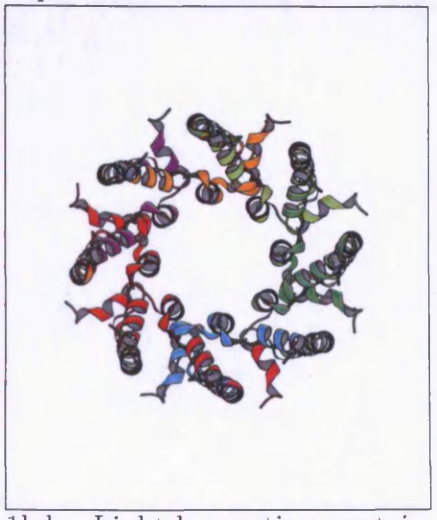
1rhz - Protein conducting channel



1kpl - CLC Cl^- channel



1okc - Adenine nucleotide carrier



1lgh - Light harvesting protein

1mxm - Mechanosensitive channel

Figure 3.13: Views of the pore-containing TM proteins along the membrane normal, showing the pore and the arrangement of TM helices. This diagram was produced using Molscript ©1997-1998, Per Kraulis. Functional pores are indicated by a pale blue circle containing a 'P'. IR indicates inward rectifying.

When modelling a channel-containing TM protein from sequence, considerable constraints could be imposed on the number of possible models if one could determine the approximate size of the pore and number of helices lining it. It is therefore significant that there is a relationship between the total number of TM helices and the number of pore-lining helices (R^2=0.67, as shown in Figure 3.14(B)). Although based on limited data, this enables the number of pore-lining helices to be easily estimated for any protein. Secondly, it is valuable to note that there is a stronger linear relationship between the number of pore-lining helices and the average pore diameter, as shown in Figure 3.14(A) (R^2=0.83). This information allows estimations to be made about the size of the pore and the number of pore-lining helices which may prove useful in modelling.

It can be seen that the diameter of the pores from the uncoupling protein TM models (see Chapter 2, Section 2.2.3) are less than the pore diameter in the dataset proteins with similar numbers of pore-lining helices. This is likely to be due to the fact that the pore diameters taken from the dataset proteins are measured from Cα to Cα of the pore-lining residues, and do not consider the positions of the residue side chains, which will tend to reduce the actual pore diameter. The data suggests that Models 1 and 3, in which 6 of the 12 TM helices line a pore of diameter 10-15Å, show the most similar arrangement of helices to that found in native proteins with 12 TM helices. Assuming they are dimeric proteins with one pore per monomer, these results therefore suggest that the UCPs are likely to show a structure more similar to Models 1 or 3 than to Model 2, in Chapter 2.

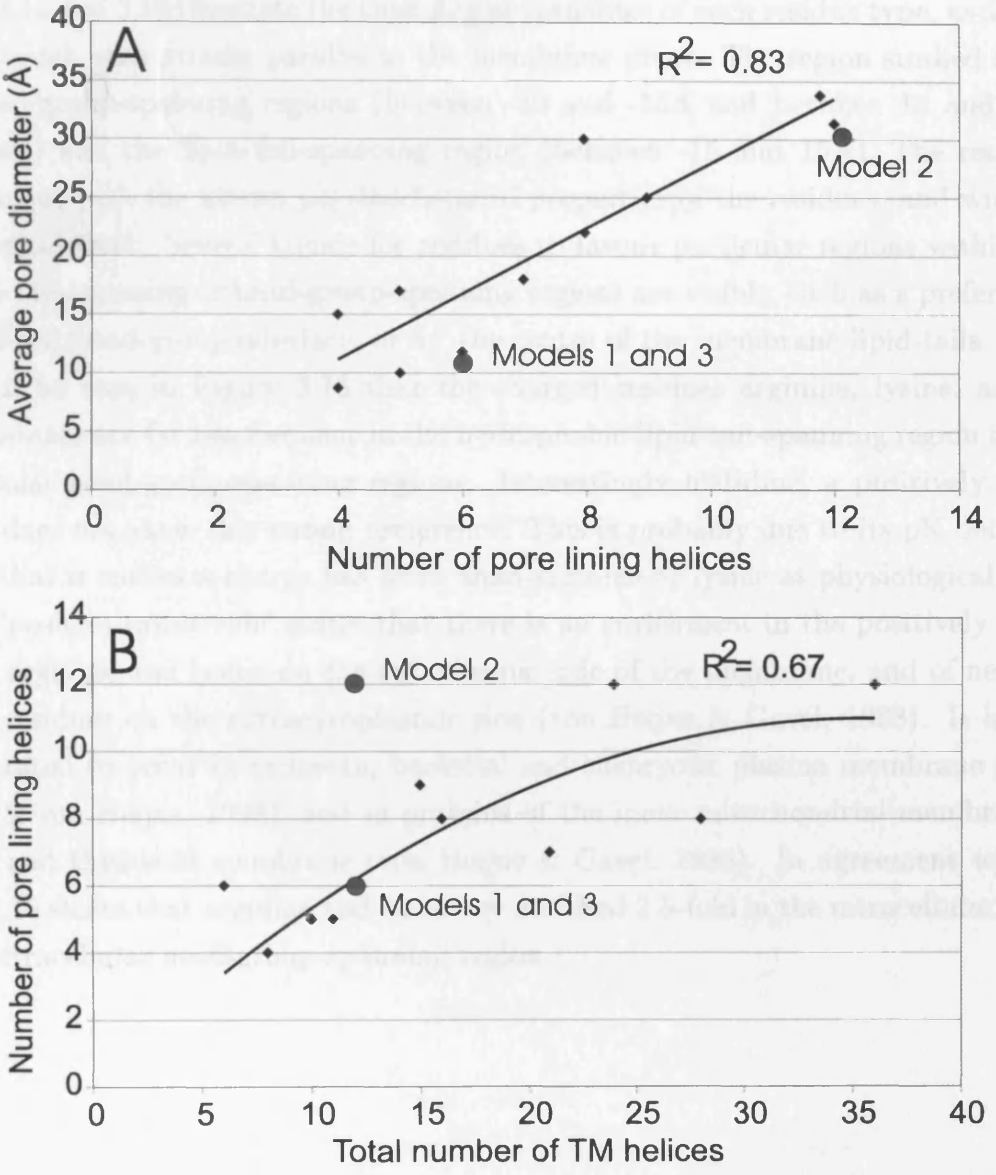


Figure 3.14: (A): Analysis of the relationship between average pore diameter and the number of pore-lining helices for both the dataset proteins (dark blue dots) and the UCP helix models described in Chapter 2 (red circles). The line shown is represented by the equation y=2.79x. (B): Analysis of the relationship between the total number of TM helices and the number of pore-lining helices for both the dataset proteins and the UCP helix models described in Chapter 2. The line shown is represented by the equation $y=0.008x^2 + 0.620x$. Regression lines and correlation coefficients were obtained using Microsoft Excel.

3.3.5 Analysis of residue propensities and hydrophobicity

3.3.5.1 Distribution of residue types across membrane-spanning regions

Figures 3.15 and 3.16 illustrate the changing propensities of each residue type, as slices are taken through each protein parallel to the membrane plane. The region studied includes both head-group-spanning regions (between -30 and -15Å and between 15 and 30Å in the figures) and the lipid-tail-spanning region (between -15 and 15Å). The results are in agreement with the known physicochemical properties of the residues, and with other experimental work. Several trends for residues to favour particular regions within either the lipid-tail-spanning or head-group-spanning regions are visible, such as a preference for the lipid-tail/head-group interface, or for the centre of the membrane lipid-tails.

It can be seen in Figure 3.15 that the charged residues arginine, lysine, aspartate and glutamate are far less frequent in the hydrophobic lipid-tail-spanning region than the highly polar head-group-spanning regions. Interestingly histidine, a positively charged residue, does not show this strong preference. This is probably due to its pK being close to 7, so that it carries a charge less often than arginine or lysine at physiological pH.

The 'positive inside rule' states that there is an enrichment in the positively charged residues arginine and lysine on the cytoplasmic side of the membrane, and of negatively charged residues on the extracytoplasmic side (von Heijne & Gavel, 1988). It has been demonstrated to occur in archaean, bacterial and eukaryotic plasma membrane proteins (Wallin & von Heijne, 1998), and in proteins of the inner mitochondrial membrane, the ER, SR and thylakoid membrane (von Heijne & Gavel, 1988). In agreement with this, Figure 3.15 shows that arginine and lysine are enriched 2.5-fold in the intracellular relative to the extracellular head-group-spanning region.

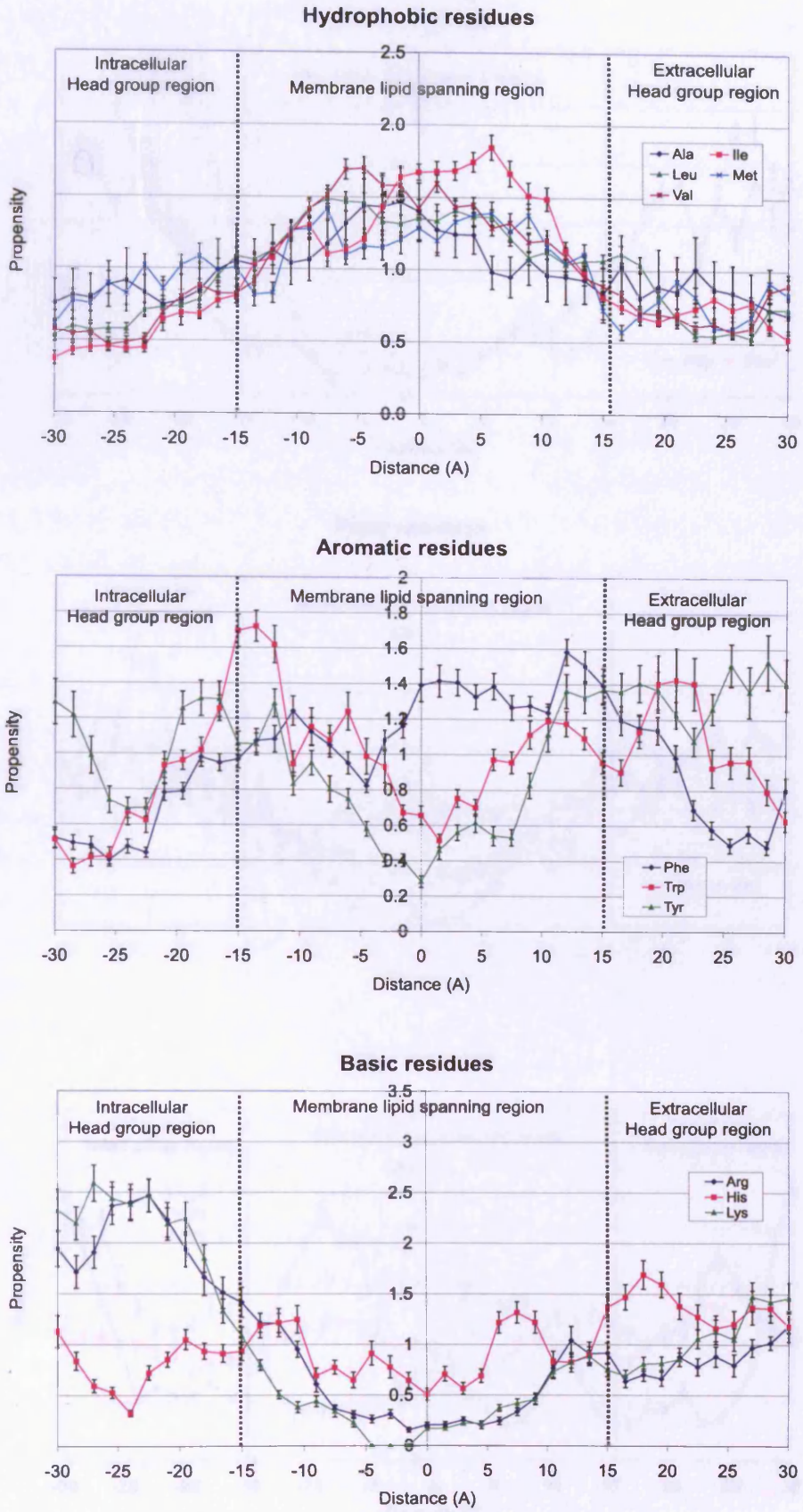


Figure 3.15: Distribution of particular residue types through the membrane-spanning region of TM proteins. The lipid-tail-spanning and head-group-spanning regions are indicated.

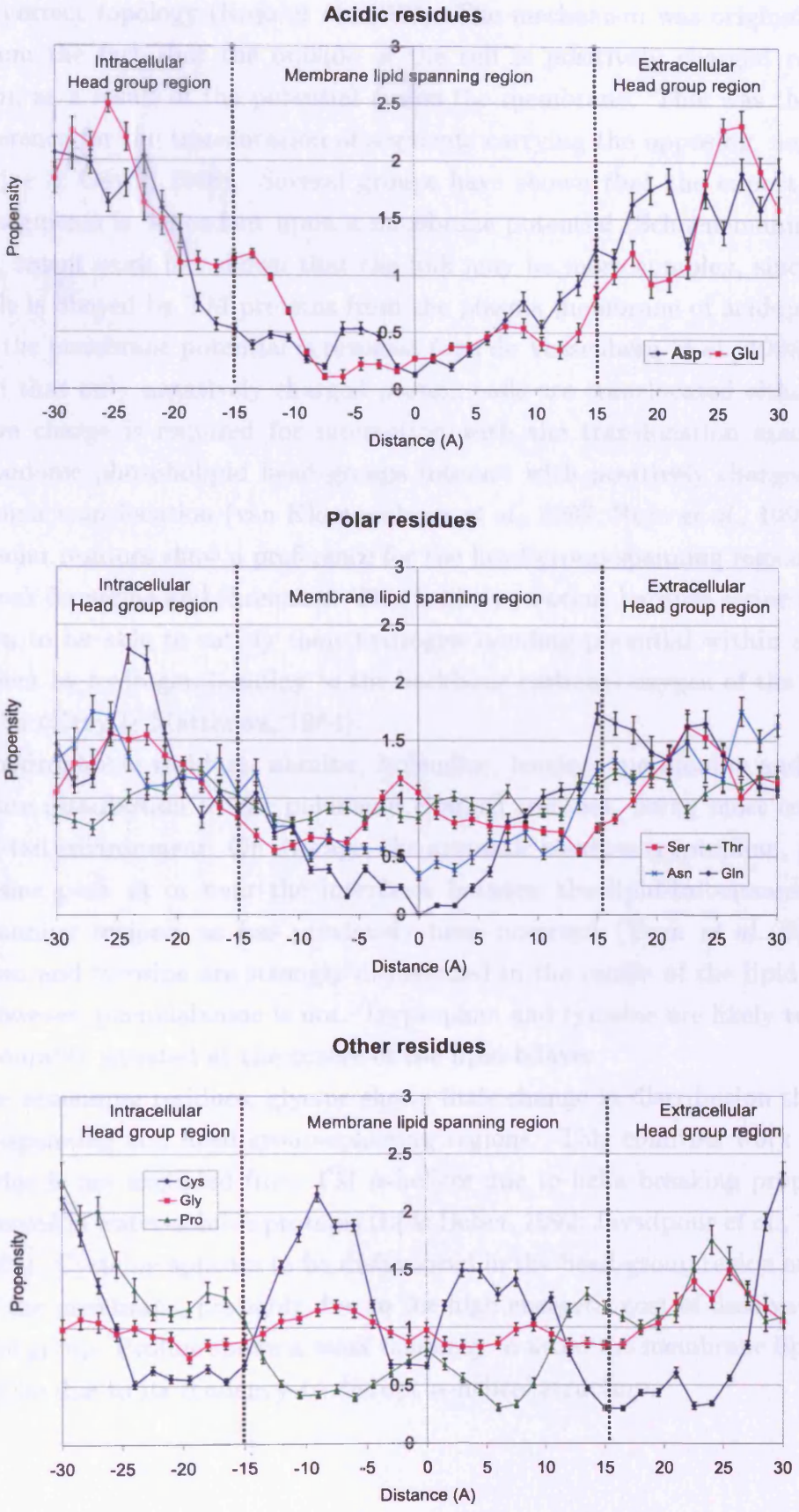


Figure 3.16: Distribution of particular residue types through the membrane-spanning region of TM proteins. The lipid-tail-spanning and head-group-spanning regions are indicated.

This charge asymmetry has been shown to play a crucial role in insertion of TM helices with the correct topology (Rojo *et al.*, 1999). The mechanism was originally thought to derive from the fact that the outside of the cell is positively charged relative to the cytoplasm, as a result of the potential across the membrane. This was thought to lead to a preference for the translocation of segments carrying the opposing, negative, charge (von Heijne & Gavel, 1988). Several groups have shown that the export of negatively charged segments is dependent upon a membrane potential (Schuenemann *et al.*, 1999). However, recent work has shown that the link may be more complex, since the positive inside rule is obeyed by TM proteins from the plasma membrane of acidophilic bacteria, in which the membrane potential is reversed (van de Vossenberg *et al.*, 1998). It has been suggested that only negatively charged protein tails are translocated either (1) because a negative charge is required for interaction with the translocation machinery or (2) because anionic phospholipid head-groups interact with positively charged regions and prevent their translocation (van Klompenburg *et al.*, 1997; Rojo *et al.*, 1999).

The polar residues show a preference for the head-group-spanning regions, although it is very weak for serine and threonine. This is likely to occur because serine and threonine are known to be able to satisfy their hydrogen bonding potential within a hydrophobic environment by hydrogen bonding to the backbone carbonyl oxygen of the previous turn in the helix (Gray & Matthews, 1984).

The hydrophobic residues, alanine, isoleucine, leucine, methionine and valine, show an opposite distribution to the polar and charged residues, being most common in the TM lipid-tail environment. On average, the aromatic residues tryptophan, phenylalanine and tyrosine peak at or near the interfaces between the lipid-tail-spanning and head-group-spanning regions, as has previously been observed (Yuen *et al.*, 2000). Whilst tryptophan and tyrosine are strongly disfavoured in the centre of the lipid-tail-spanning region, however, phenylalanine is not. Tryptophan and tyrosine are likely to be too polar to be favourably situated at the centre of the lipid bilayer.

Of the remaining residues, glycine shows little change in distribution throughout the lipid-tail-spanning and head-group-spanning regions. This confirms work that suggests that glycine is not excluded from TM α-helices due to helix-breaking properties, as has been observed in water-soluble proteins (Li & Deber, 1992; Javadpour *et al.*, 1999; Bywater *et al.*, 2001). Cysteine appears to be disfavoured in the head-group region and at the very centre of the membrane, probably due to the high energetic cost of desolvating its highly polar thiol group. Proline shows a weak tendency to avoid the membrane lipid-tail region. This may be due to its tendency to disrupt α-helical structure.

3.3.5.2 Comparison of the residue composition of the lipid-tail-spanning and head-group-spanning regions

As would be expected, the regions of protein accessible to lipid-tails and head-groups show characteristic differences in their amino acid composition. These are shown in the propensities in Table 3.5 and Figure 3.17. Whilst these propensities give a useful estimate of the preferences of different amino acids for either the head-group or lipid-tail region, it should be emphasised that much data is lost by this representation. Single-value propensities imply that the occurrence of a residue is constant throughout the entire lipid-tail-spanning or head-group-spanning region, whilst in fact there is much variation. This variation is derived from the differing environments experienced by residues of both differing lipid-tail-accessibility, (as discussed in Section 3.3.5.4), and at differing heights within the membrane, (as shown in Figures 3.15 and 3.16). It seems likely that, due to this heterogeneity within each region, position-specific propensities would be more effective for predictive purposes.

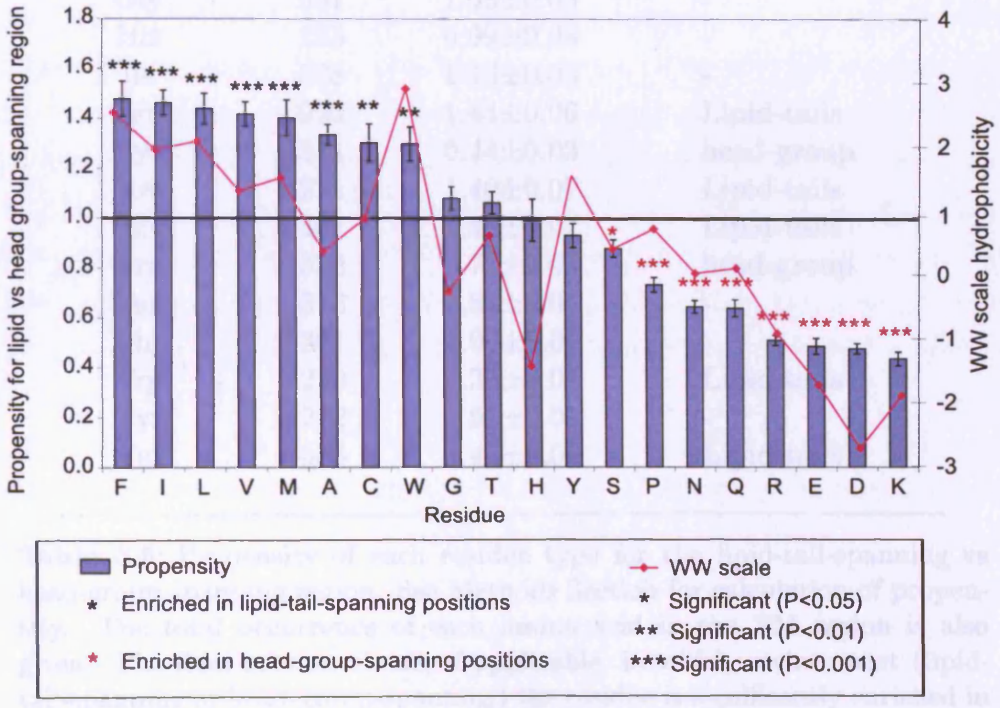


Figure 3.17: Comparison of the amino acid composition of the lipid-tail-spanning and head-group-spanning regions of TM proteins, indicating the high degree of correlation with the WW hydrophobicity scale. A propensity of greater than 1 indicates an enrichment in the lipid-tail-spanning region relative to the head-group-spanning region. Similarly, a propensity of less than 1 indicates an enrichment in the head-group-spanning region relative to the lipid-tail-spanning region.

Residue	**Occurrence**	**Propensity**	**Significantly enriched in**
Ala	724	1.33±0.04	Lipid-tails
Arg	248	0.51±0.02	head-group
Asn	228	0.64±0.02	head-group
Asp	193	0.48±0.02	head-group
Cys	74	1.30±0.08	Lipid-tails
Gln	176	0.64±0.03	head-group
Glu	208	0.49±0.03	head-group
Gly	591	1.08±0.05	-
His	185	0.99±0.08	-
Ile	538	1.46±0.05	-
Leu	920	1.44±0.06	Lipid-tails
Lys	211	0.44±0.03	head-group
Met	221	1.40±0.07	Lipid-tails
Phe	461	1.48±0.07	Lipid-tails
Pro	332	0.73±0.03	head-group
Ser	376	0.88±0.03	-
Thr	361	1.06±0.04	-
Trp	230	1.30±0.07	Lipid-tails
Tyr	222	0.93±0.05	-
Val	583	1.42±0.05	Lipid-tails

Table 3.5: Propensity of each residue type for the lipid-tail-spanning vs head-group-spanning region. See Methods Section for calculation of propensity. The total occurrence of each amino acid in the TM region is also given. The final column states, if applicable, in which environment (lipid-tail-spanning or head-group-spanning) the residue is significantly enriched in comparison with the other environment ($P < 0.05$).

As can be seen from the propensities in Table 3.5 and Figure 3.17, the lipid-tail-spanning region, (as defined in Figure 3.1), is enriched in the hydrophobic and aromatic amino acids (W, F, L, I, M, V and A), relative to the head-group region. In contrast, the head-group region is enriched in polar and charged residues, (D, E, Q, K, R and N). Proline, whilst being relatively hydrophobic, is strongly enriched in the head group region relative to the lipid-tail-spanning region. This may be due to its ability to kink or break helices, in order to terminate the TM helices on either side of the membrane. Cysteine is strongly enriched in the lipid-tail-spanning region, for reasons that are not understood. Interestingly, the polar residues serine and threonine show very little preference for either location, suggesting that they can be well tolerated, or have valuable functions, in both environments. This is likely to be due to their ability to hydrogen bond with solvent, as in the head-group region, and with the previous backbone carbonyl oxygen of the same chain in a hydrophobic lipid-tail environment (Gray & Matthews, 1984).

Figure 3.17 illustrates the high degree of correlation between the propensities and White and Wimley scale (Wimley *et al.*, 1996; Jayasinghe *et al.*, 2001) hydrophobicity, emphasising the trend that the greater the hydrophobicity of a particular residue the greater its preference for the lipid-tail-spanning region. The 'AVILM content' of a particular region was calculated as the percentage of the total residues that were of type A, V, I, L, or M, and displayed in Figure 3.18. This figure illustrates the strength of the preference of hydrophobic residues for lipid-tail-spanning regions over head-group spanning regions. Two clearly separate distributions can be seen, indicating that lipid-tail-spanning regions tend to have a greater proportion of hydrophobic residues than the head-group-spanning region. The average AVILM content of lipid-tail-spanning regions is 20% higher than that in head-group-spanning regions.

The distributions of AVILM content for both lipid-tail-spanning and head-group-spanning residues have a range of approximately 30%. This range is unlikely to indicate a possible variation in the methods by which different proteins achieve stability within the membrane environment. Instead it suggests that it is the relative difference in hydrophobicity between TM and non-TM segments that drives insertion, rather than the absolute values.

3.3.5.3 Comparison of the hydrophobicity of lipid-tail-accessible and buried lipid-tail-spanning regions

Whilst a similar separation was expected for buried residues versus lipid-tail-accessible residues within the lipid-tail-spanning region, this was not observed, as is shown in Figure 3.19. There is no significant difference between the total AVILM content of buried and

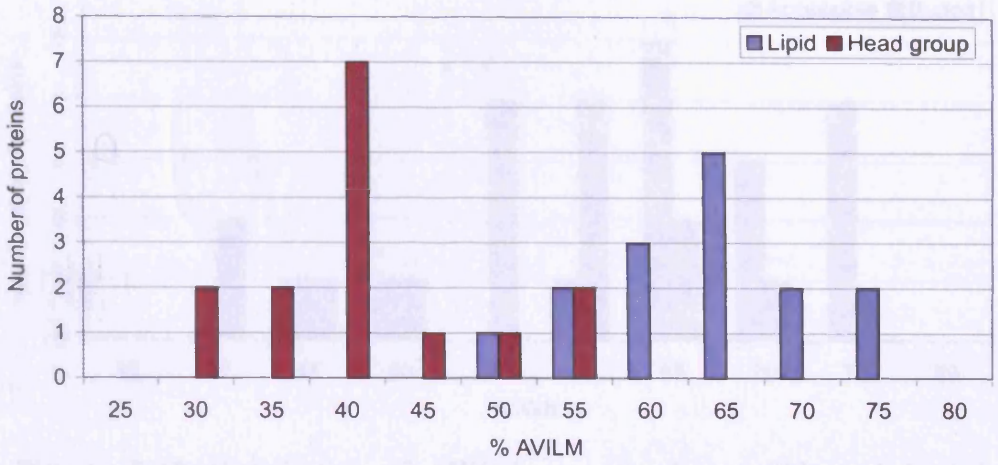


Figure 3.18: Distribution of the AVILM content of TM regions spanning the lipid-tail region and head-group region of the bilayer for each dataset protein. Two clearly separate distributions can be seen, indicating that lipid-tail-spanning regions tend to have a greater proportion of hydrophobic residues than the head-group-spanning region. AVILM content was calculated as the percentage of the total residues that were of type A, V, I, L, or M.

lipid-tail-accessible regions (mean percentages of residues AVILM are 48% in buried and 53% in lipid-tail-accessible positions, P = 0.148). Both buried and lipid-tail-accessible residues within the lipid-tail region showed a great range of AVILM content, from 35 to 70% and 40 to 75% respectively. To account for the huge variation, there may be differing driving forces that have caused particular proteins to adopt these different proportions of hydrophobic residues in buried and lipid-tail-accessible positions. However, these forces do not appear to be linked to the role of the protein, since there is no significant difference between the proportions of hydrophobic, charged, aromatic or polar residues in buried or lipid-tail-accessible positions of proteins of different function (ion pumps, ion channels, electron carriers or miscellaneous; data not shown). Again, this suggests that it is the relative difference in hydrophobicity between groups of residues, (both between lipid-tail-spanning and head-group-spanning residues and between buried and lipid-tail-accessible residues), that promotes TM protein folding and membrane insertion, rather than the absolute values. Further work, particularly more TM protein structures, will be needed before the reasons for this variation in hydrophobic amino acid content can be fully understood.

However, if we consider each helix separately, we find that lipid-tail-accessible residues are significantly more hydrophobic, according to the WW hydrophobicity scale, than

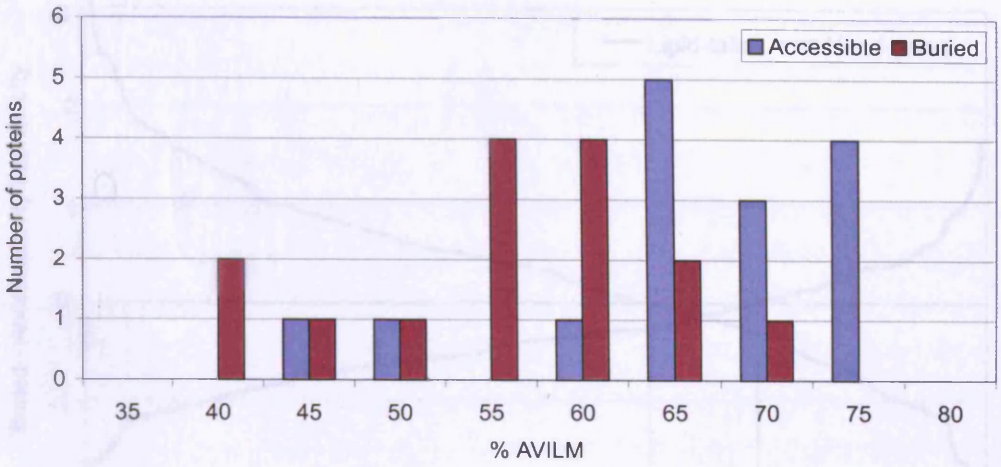


Figure 3.19: Distribution of AVILM content of accessible and buried residues within the lipid-tail-spanning region. Distinct distributions cannot be seen, indicating that accessible and buried residues have similar proportions of hydrophobic residues. AVILM content was calculated as the percentage of the total residues that were of type A, V, I, L, or M.

buried ones. (The mean hydrophobicity scores are -0.47 for 2510 lipid-tail-accessible residues and -0.12 for 1528 buried residues, $P < 0.0001$). This is illustrated in Figure 3.20. A similar result was achieved using the Kyte and Doolittle scale (Kyte & Doolittle, 1982) (data not shown). As expected, the reverse trend is observed in the head-group-spanning region, where the buried residues are more hydrophobic that the accessible ones ($p<0.001$). This is shown in Figure 3.20.

As shown in Figure 3.20 that far less discrimination between buried and accessible residues can be achieved using hydrophobicity data from the head-group region than from the lipid-tail-spanning region. In fact it seems that there is an inherent level of hydrophobicity found in the core of all protein regions: in water-soluble proteins, in TM lipid-tail-spanning regions and in TM head-group-spanning regions. (The similar level of hydrophobicity of lipid-tail-spanning and head-group- spanning residues can be seen in Figure 3.20. That the core region of water-soluble proteins is of similar hydrophobicity to that of the lipid-tail region has been shown by other groups (Chothia, 1976)). Interestingly, this similarity in hydrophobicity implies that the packing interactions that occur in each of these environments are likely to be similar. Whether this is the case will be investigated in the following sections.

Whilst the core hydrophobicity of all proteins appears to remain relatively constant, the hydrophobicity of the surface residues is modified in order to facilitate interaction with the external environment: either water, membrane lipid-tails or membrane head-groups.

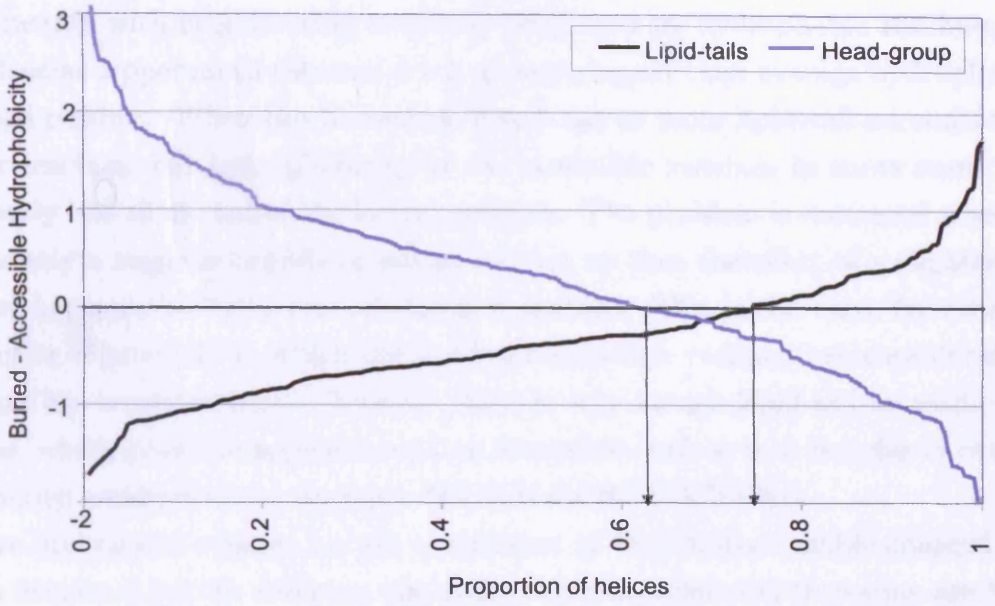


Figure 3.20: Comparison of the hydrophobicity, according to the White and Wimley scale, of lipid-tail/head-group-accessible and buried residues for each of the dataset TM helices. Plotted is the difference in hydrophobicity between accessible and buried residues on each helix. Only residues with Cα atoms in the lipid-tail-spanning region or head-group-spanning region of each helix are analysed. Homologous chains and segments lacking either accessible or buried residues were removed.

In the case of head-group-spanning residues, however, the inherent hydrophobicity of the buried residues appears to be similar to that required for optimal interaction with the phospholipid head-groups. This results in a very small discrimination between the hydrophobicity of head-group-accessible and buried residues in this environment. As a result, prediction of buried and accessible residues, at least using differences in hydrophobicity, is likely to be much more effective within the lipid-tail-spanning region than the head-group region.

For predictive power we need to consider not just mean hydrophobicity values but also the strength of the relationship and the frequency with which exceptions occur. We found that in 74% of helices the mean hydrophobicity is greater for lipid-tail-accessible residues than for buried residues. Therefore, hydrophobicity can make a valuable contribution to prediction but, since the theoretical maximum accuracy of prediction is only 74%, other parameters will be required to increase the performance of the prediction.

Conversely, 26% of helices do not show the expected hydrophobicity characteristics, and some have lipid-tail-accessible residues that are considerably less hydrophobic than

their buried residues (Figure 3.20). The situation seems to occur mainly in helices where the interaction with neighbouring helices is performed by hydrophobic residues, such as via the leucine zipper motif (Section 3.1.2, giving a higher than average hydrophobicity of the buried residues. When this is combined with one or more lipid-tail-accessible charged or polar residues, the hydrophobicity of the accessible residues in some cases becomes significantly less than that of the buried residues. The problem is increased when a helix contains only a single accessible or buried residue, so that the effect of a slightly unusual degree of hydrophobicity in this residue is magnified. This is the case, for example, for one helix in Figure 3.20 in which the lipid-tail-accessible residues are considerably more polar than the buried residues. However, there is only a single lipid-tail-accessible residue, a glycine, which gives the appearance of an accessible surface that is polar in comparison to the buried residues in the sequence HHAIALGLHTTTLILVKG.

There are various reasons for the occurrence of lipid-tail-accessible charged residues given in Section 3.3.8. In addition, the polar residues serine and threonine are known to be more commonly accessible to lipid-tails than predicted from their polarity, due to their ability to form hydrogen bonds with the preceding turn of the helix (Gray & Matthews, 1984). These findings confirm that the folding of TM proteins does not simply rely on the principle of burying all hydrophilic residues and exposing all hydrophobic ones.

As described previously, there is no significant difference between the total AVILM content of buried and lipid-tail-accessible regions. This suggests that, while the total numbers of hydrophobic residues in each environment are similar, the proportions of each individual residue must be different, leading to the observed difference in WW hydrophobicity score. Hence it is not simply being classified as a hydrophobic amino acid that leads to a preference for lipid-tail-accessible positions, but the specific physico-chemical properties of each amino acid, or the degree of hydrophobicity. In agreement with this, certain individual amino acids do show a significant preference for either accessible or buried environments, as described in the next section (Section 3.3.5.4).

In the head-group-spanning region, only 61% of helices show the expected trend, with buried residues more hydrophobic than lipid-tail-accessible residues. This much weaker relationship suggests that TM helix packing methods that make use of hydrophobicity data should probably consider the lipid-tail-spanning residues alone.

3.3.5.4 Comparison of the preferences of particular residues for lipid-tail-accessible vs buried lipid-tail-spanning regions

Within the lipid-tail-spanning region, certain amino acids show a preference for either accessible or buried environments, as shown in Figure 3.21 and Table 3.6. This preference

can be quantified for each residue using a propensity value (see Section 3.2.4.10). The hydrophobic amino acids leucine, phenylalanine and tryptophan both show a preference for lipid-tail-accessible positions. Consistent with the work of Ulmschneider & Sansom (2001), alanine shows a preference for buried positions, but the other hydrophobic residues show no significant preference. This is at first sight somewhat unexpected, given the hydrophobicity of the bilayer. However, it appears to be due to a genuine lack of preference, with propensity values very close to 1, rather than a lack of statistical power due to the size of the dataset. It probably reflects both the ability of hydrophobic residues to interact favourably with the membrane lipid-tails and their important role in leucine zipper packing (Crick, 1953; Langosch & Heringa, 1998) of the TM helices.

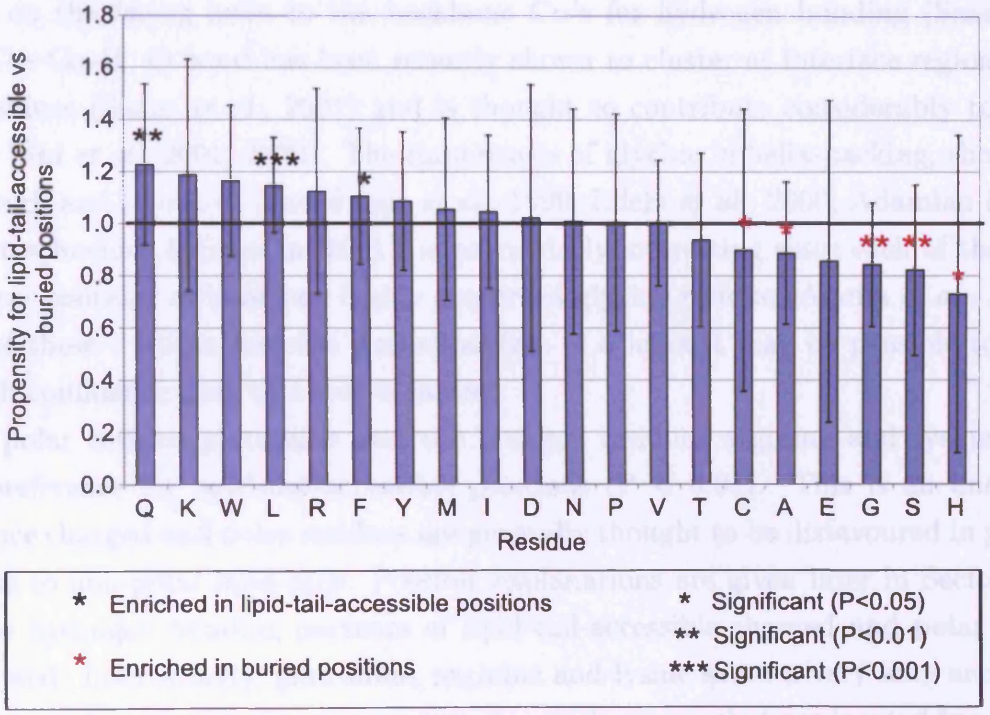


Figure 3.21: Comparison of the amino acid composition of the buried and accessible membrane lipid-tail-spanning residues of 24 TM proteins. A value of greater than 1 indicates an enrichment in the lipid-tail-accessible positions relative to buried positions. Similarly, a value of less than 1 indicates an enrichment in the buried positions relative to lipid-tail-accessible positions.

In general, lipid-tail-accessible positions are slightly enriched in aromatic residues compared to buried ($P < 0.05$). This may be due to the fact that such large aromatic residues are difficult to pack efficiently between helices. In addition, aromatic residues are thought to play an important role in the anchoring of TM proteins at the correct height within

the bilayer (Yuen *et al.*, 2000), for which an accessible position is likely to be required.

Glycine, alanine, histidine, cysteine and serine show strong preferences for buried positions. These results are consistent with the recent work of several groups who have shown the importance of glycine, alanine, serine and threonine in the close packing of TM helix interfaces (Ulmschneider & Sansom, 2001; Javadpour *et al.*, 1999; Eilers *et al.*, 2000; Senes *et al.*, 2000). Threonine shows a similar, but weaker and non-significant, preference. Overall, as would be expected, buried positions were enriched in polar residues compared to lipid-tail-accessible ones (mean percentage of residues that are polar in lipid-tail-accessible positions is 9% and in buried positions is 15%, $P = 0.01$).

The interaction of these small and polar residues forms the second major method by which TM helices pack, a mechanism that is rarely observed in water-soluble proteins (Eilers *et al.*, 2002). Their small side chain volumes allow the close approach of polar residues on the facing helix to the backbone Cα's for hydrogen bonding (Senes *et al.*, 2001). The Cα-H$\cdots$O bond has been recently shown to cluster at interface regions rich in these residues (Senes *et al.*, 2001) and is thought to contribute considerably to protein stability (Shi *et al.*, 2001, 2002). The importance of glycine in helix-packing, shown both in this work and in others (Javadpour *et al.*, 1999; Eilers *et al.*, 2000; Adamian & Liang, 2001; Ulmschneider & Sansom, 2001), is particularly interesting since each of the 6 UCP TM helices contains at least one highly conserved glycine residue (Aquila *et al.*, 1985). If several of these residues line one particular face of a helix it may be possible to predict with high confidence that this face is buried.

The polar residue glutamine and the charged residues arginine and lysine show a strong preference for lipid-tail-accessible positions ($P < 0.01$). This is an unexpected result since charged and polar residues are generally thought to be disfavoured in positions accessible to non-polar lipid-tails. Possible explanations are given later in Section 3.3.8, when the hydrogen bonding partners of lipid-tail-accessible charged and polar residues are analysed. Interestingly, glutamine, arginine and lysine share a very long and flexible side chain, perhaps facilitating their interaction with groups that are located too far away for other residues to reach. The other charged residues, with shorter side chains, tend to be buried (glutamate) or more evenly distributed between buried and accessible positions (aspartate).

3.3.5.5 The lipid-tail-accessibility scale

There is very little correlation between the traditional hydrophobicity scales and the propensity of residues to be lipid-tail-accessible (Figure 3.22). Hence, a scale that can accurately predict the likely environment (buried or lipid-tail-accessible) of a lipid-tail-

spanning residue must take into account these deviations between the two measures. A lipid-tail-accessibility (LA) scale was derived for this purpose from the observed propensities of lipid-tail-spanning residues to be buried.

Residue	**Occurrence**	**Propensity**	**Significantly enriched in**
Ala	555	0.89±0.27	Buried
Arg	100	1.12±0.40	-
Asn	93	1.01±0.43	-
Asp	72	1.02±0.51	-
Cys	63	0.90±0.54	Buried
Gln	83	1.23±0.31	Accessible
Glu	90	0.86±0.62	-
Gly	411	0.84±0.24	Buried
His	91	0.73±0.61	Buried
Ile	492	1.05±0.29	-
Leu	771	1.15±0.18	Accessible
Lys	90	1.19±0.44	-
Met	183	1.06±0.35	-
Phe	434	1.11±0.26	Accessible
Pro	162	1.00±0.42	-
Ser	203	0.82±0.33	Buried
Thr	255	0.94±0.33	-
Trp	154	1.16±0.29	-
Tyr	92	144±0.27	-
Val	510	1.00±0.24	-

Table 3.6: Propensity of each residue type to be found in accessible vs buried positions in the lipid-tail-spanning region. The total occurrence of each amino acid in the TM-lipid-tail-spanning region is also given. The final column states, if applicable, in which environment (lipid-tail-accessible or buried) the residue is significantly enriched in comparison with the other environment ($P < 0.05$).

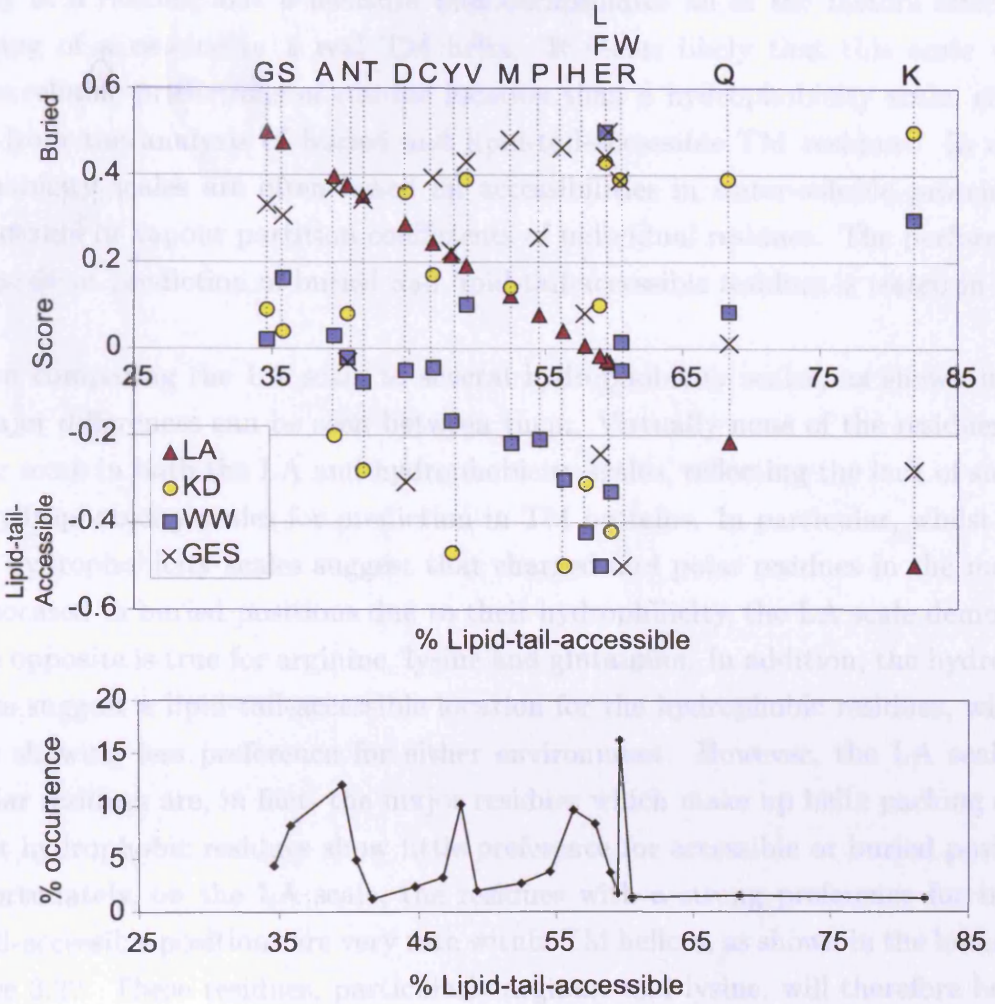


Figure 3.22: Upper panel: Plot of the White and Wimley (WW) hydrophobicity scale (Wimley *et al.*, 1996; Jayasinghe *et al.*, 2001), the Kyte and Doolittle (KD) hydrophobicity scale (Kyte & Doolittle, 1982), the GES scale (Engelman *et al.*, 1986) and our lipid-tail-accessibility (LA) scale against the accessible propensity of each residue. The LA scale is derived from this residue propensity data. The correlation coefficient (R^2) for each scale with the propensity data is 1 for the LA scale, 0.07 for the KD scale, 0.002 for the WW scale and 0.14 for the GES scale. The values for a single residue on each scale are joined by a dotted line. Lower panel: % occurrence of each residue in the lipid-tail-spanning region.

The value that a residue receives in the LA scale, shown in Figure 3.22, is simply computed by proportionally scaling the residue propensities into the desired range of -0.5 to +0.5. Hence, this is not a traditional hydrophobicity scale, representing the water/lipid solubility of a residue, but a measure that encapsulates all of the factors affecting the positioning of a residue in a real TM helix. It seems likely that this scale will give far more reliable predictions of residue location than a hydrophobicity scale, since it is derived from the analysis of buried and lipid-tail-accessible TM residues. In contrast, hydrophobicity scales are often based on accessibilities in water-soluble proteins or on water/octanol or vapour partition coefficients of individual residues. The performance of the LA scale at prediction of buried and lipid-tail-accessible residues is tested in Chapter 4.

When comparing the LA scale to several hydrophobicity scales, as shown in Figure 3.22, major differences can be seen between them. Virtually none of the residues receive a similar score in both the LA and hydrophobicity scales, reflecting the lack of suitability of the hydrophobicity scales for prediction in TM proteins. In particular, whilst the traditional hydrophobicity scales suggest that charged and polar residues in the membrane will be located in buried positions due to their hydrophilicity, the LA scale demonstrates that the opposite is true for arginine, lysine and glutamine. In addition, the hydrophobicity scales suggest a lipid-tail-accessible location for the hydrophobic residues, with polar residues showing less preference for either environment. However, the LA scale shows that polar residues are, in fact, the major residues which make up helix packing contacts and that hydrophobic residues show little preference for accessible or buried positions.

Unfortunately, on the LA scale, the residues with a strong preference for buried or lipid-tail-accessible positions are very rare within TM helices, as shown in the lower section of Figure 3.22. These residues, particularly arginine and lysine, will therefore be unable to make a major contribution to the prediction of helix face accessibility. Conversely, the very common residues, such as leucine and isoleucine, show very little preference for either environment. It is likely that this effect will be the limiting factor in the predictive power of the scale.

3.3.6 Analysis of residue sequence conservation

As shown in Figure 3.23, lipid-tail-accessible residues are significantly less conserved in terms of sequence than buried residues in the lipid-tail-spanning region. (Mean conservation scores calculated by SCORECONS (Section 3.2.4.2) are 0.63 and 0.68 respectively, $P < 0.001$, for 2115 lipid-tail-accessible and 1231 buried residues). Similarly, buried residues also appear to be more conserved than accessible ones throughout the whole

protein (i.e. including both membrane-accessible and solvent-accessible residues). This is probably to be due to the requirement of buried residues to be conserved in order to maintain favourable interactions with neighbouring residues. Accessible residues, on the other hand, are under only weak selective pressure to be conserved, in order to facilitate more general interactions with solvent or lipid-tails.

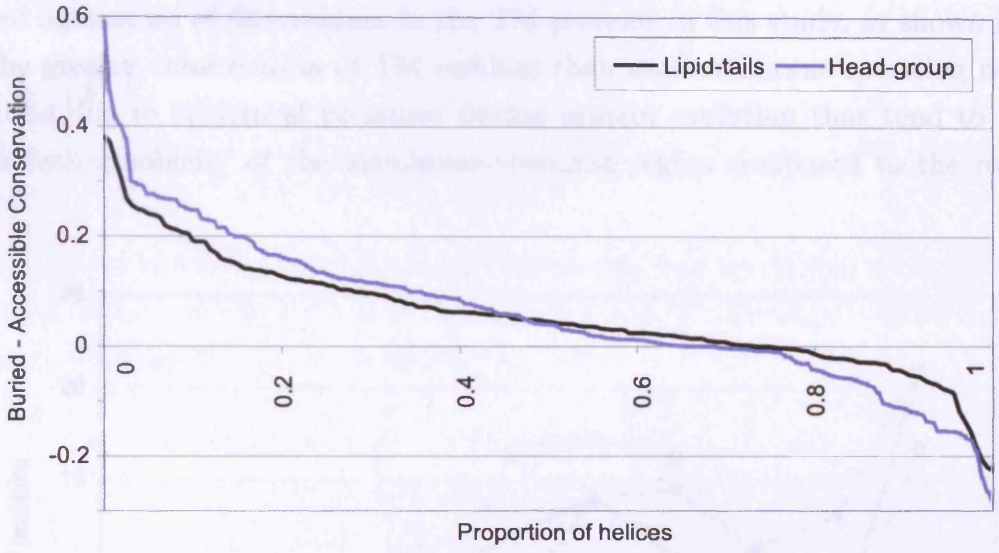


Figure 3.23: Comparison of the conservation scores of lipid-tail/head-group-accessible and buried residues for each of the dataset TM helices. Plotted is the difference in hydrophobicity between accessible and buried residues on each helix. Only residues with Cα atoms in the lipid-tail-spanning region or head-group-spanning region of each helix are analysed. Homologous chains, segments lacking either accessible or buried residues, or those for which no homologous sequences could be obtained, were removed.

Whilst there is great variability in the actual average conservation score of buried and lipid-tail-accessible residues between proteins, and even between helices in the same protein, in 73% of helices the buried positions are more conserved than the lipid-tail-accessible ones. A slightly weaker trend is observed when comparing the sequence conservation of accessible and buried head-group-spanning residues (shown in Figure 3.23), where 65% of helices have buried residues more conserved than head-group-accessible residues.

These results suggest that conservation is a method by which the buried face of many TM helices could be identified, although other information would be needed to improve the accuracy above the maximum of 73%. In addition, similar to the results for hydrophobicity, the conservation of residues within the lipid-tail-spanning region is likely to make a slightly more reliable predictor of helix packing than in the head-group-spanning region.

64% of helices showed lipid-tail-accessible residues that are both less conservation and

more hydrophobic than the buried residues. The variability between proteins suggests that in model prediction it will not be possible to identify lipid-tail-accessible residues simply based upon a cut-off of conservation score. However, the results suggest that, on comparing the conservation or hydrophobicity of each face of the helix, the most conserved or polar face is very likely to be buried.

The lipid-tail-spanning residues are more conserved than residues in general, when compared against all of the residues in the TM proteins in this study, as shown in Figure 3.24. The greater conservation of TM residues than non-membrane-spanning residues is likely to be due to structural pressures during protein evolution that tend to maintain the high hydrophobicity of the membrane-spanning region compared to the rest of the protein.

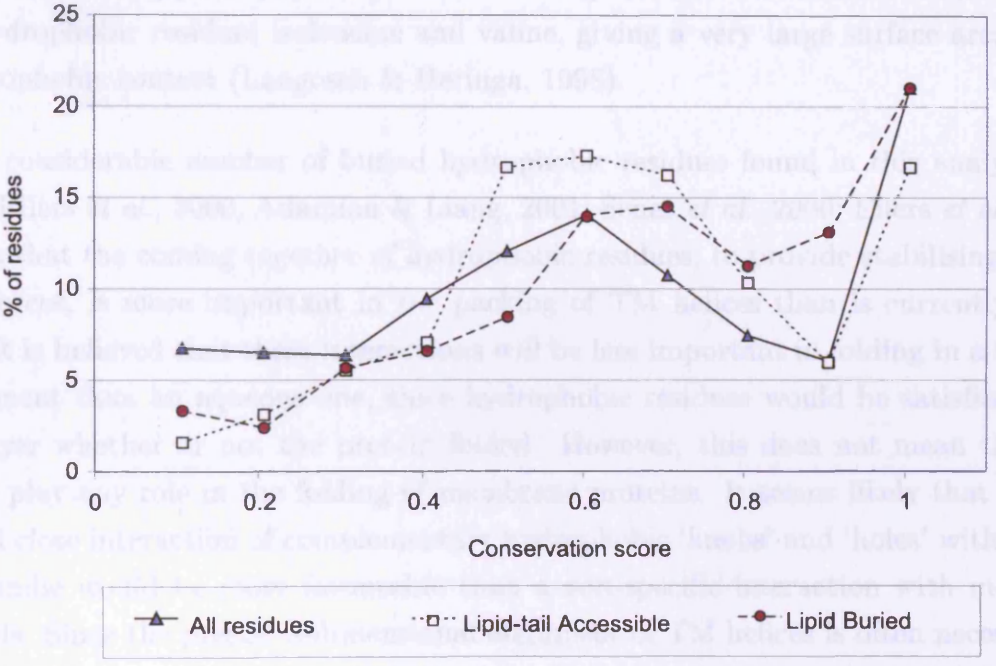


Figure 3.24: Comparison of the distributions of conservation scores for lipid-tail-spanning residues and all residues in 24 TM proteins. 'All residues' includes both membrane- and non-membrane-spanning residues.

3.3.7 Role of buried hydrophobic residues in transmembrane helix packing

It seems that the rules governing the locations of residues are more complex than the simple assumptions that all hydrophobic residues will prefer to be accessible to lipid-tails

and all charged and polar residues will prefer buried environments. It therefore seems likely that problems that arise in hydrophobic moment-based prediction methods may be due to a number of residue types that do not conform to their expected environment preferences. Hydrophobic residues, for example, are common at interfaces for two reasons:

1. The small side chain volumes of alanine and glycine allow the close approach of polar residues on the facing helix, such as serine and threonine, to the backbone Cα's for hydrogen bonding (Senes *et al.*, 2001). Hence, despite their relative hydrophobicity, the Cα-H$\cdots$O bond has been recently shown to cluster at interface regions rich in these residues (Senes *et al.*, 2001) and is thought to contribute considerably to protein stability (Shi *et al.*, 2001, 2002).

2. Alanine and glycine provide pockets or troughs, into which can fit the β-branched hydrophobic residues isoleucine and valine, giving a very large surface area for hydrophobic contact (Langosch & Heringa, 1998).

The considerable number of buried hydrophobic residues found in this analysis, and others (Eilers *et al.*, 2000; Adamian & Liang, 2001; Senes *et al.*, 2000; Eilers *et al.*, 2002), suggests that the coming together of hydrophobic residues, to provide stabilising van der Waals forces, is more important in the packing of TM helices than is currently understood. It is believed that these interactions will be less important in folding in a lipid-tail environment than an aqueous one, since hydrophobic residues would be satisfied within the bilayer whether or not the protein folded. However, this does not mean that they will not play any role in the folding of membrane proteins. It seems likely that the specific and close interaction of complementary hydrophobic 'knobs' and 'holes' within a TM helix bundle would be more favourable than a non-specific interaction with membrane lipid-tails. Since the precise 3-dimensional alignment of TM helices is often necessary for function, interactions of large scale hydrophobic regions may be more useful than pairing of many buried charges, which may result in the protein being too polar for membrane insertion.

3.3.8 Proposed roles for lipid-tail-accessible charged residues

There are several possible reasons why charged residues might be found in positions classified in this study as lipid-tail-accessible. These are:

- The charged residues may be paired
- The charged residues may be near the interface with the head-group region

- The charged residues may line a pore or water-filled cavity and therefore not be truly lipid-tail-accessible
- The charged residues may be hydrogen bonded to other residues, water or cofactors

The contributions of each of these reasons to the number of lipid-tail-accessible charged residues was assessed, as shown in Figure 3.25(A). For comparison, the calculations were repeated for lipid-tail-accessible polar residues (Figure 3.25(B)). The results of these analyses are discussed in this section.

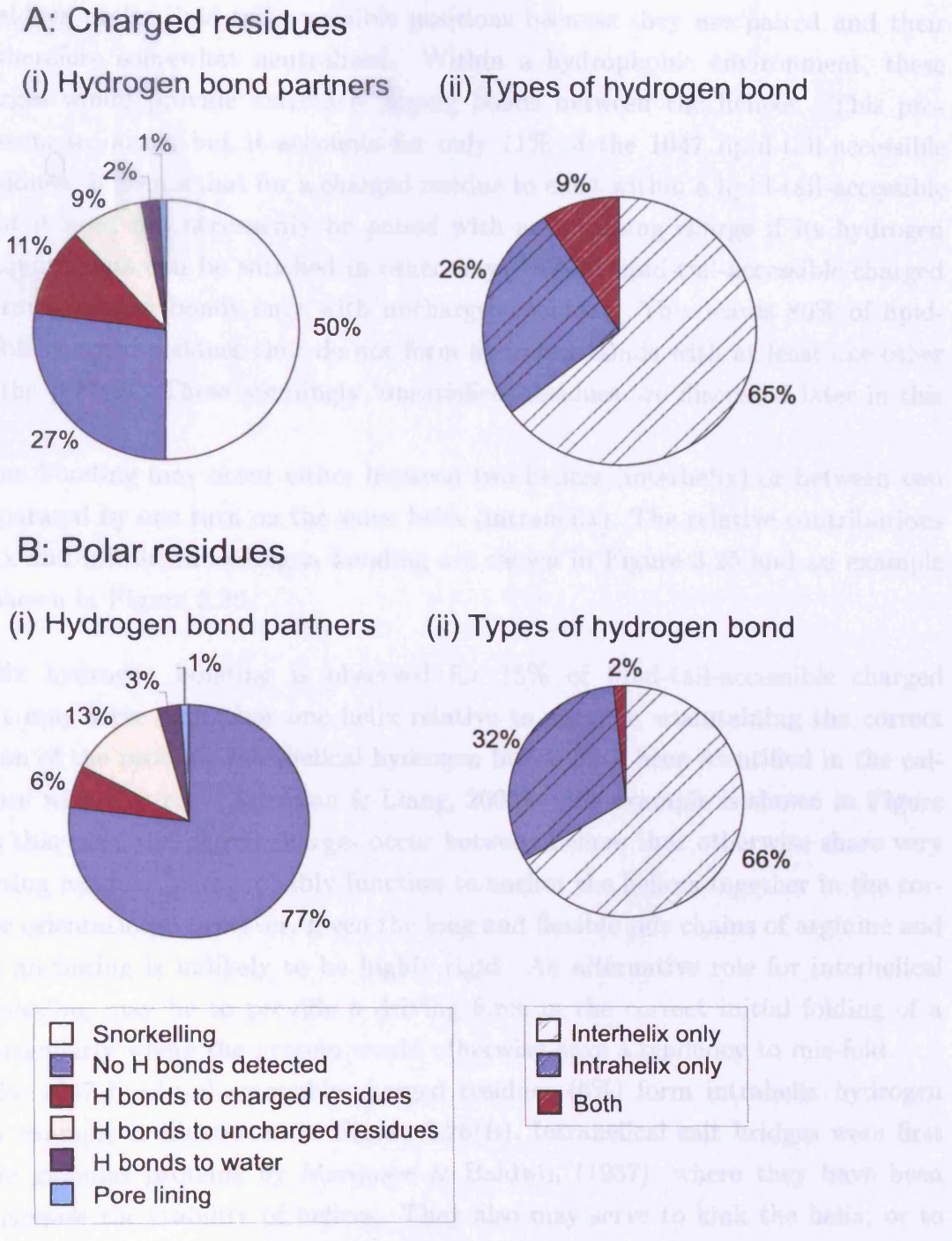


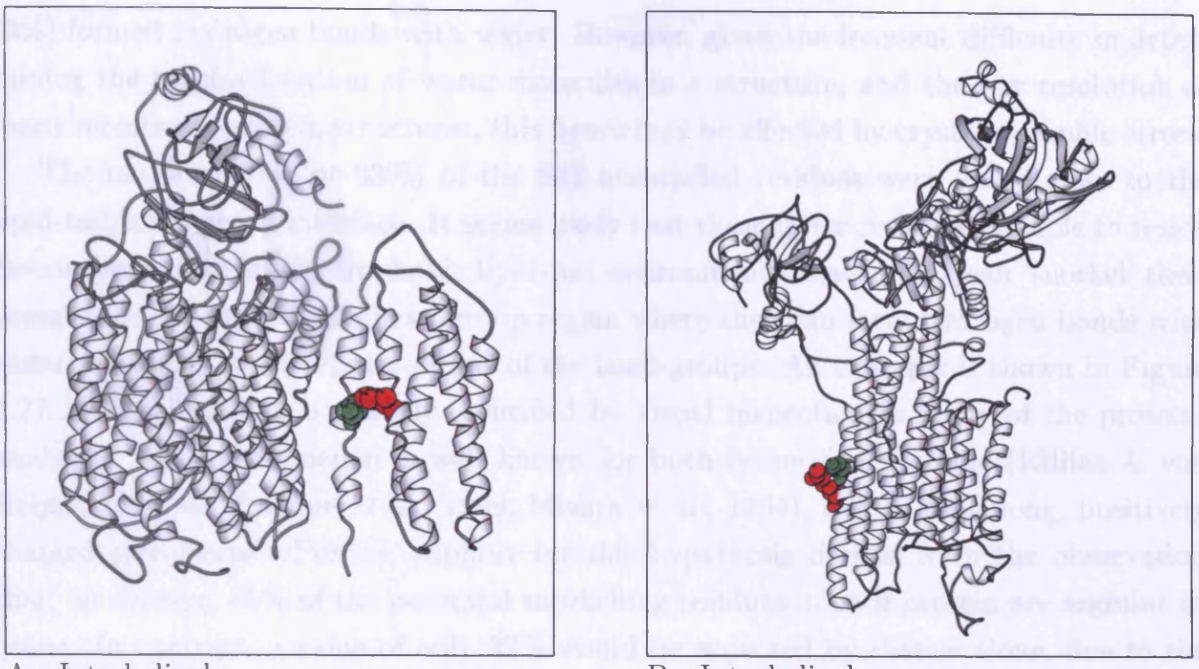
Figure 3.25: (i) Hydrogen (H) bonding partners and (ii) types of hydrogen bonds for (A) the 1047 observed lipid-tail-accessible charged residues and (B) the 1202 lipid-tail-accessible polar residues in the dataset. Hydrogen bonds were detected by HBPlus v 3.0 (McDonald & Thornton, 1994). Main-chain/main-chain hydrogen bonds are excluded. Since both hydrogen bonding partners are included, each hydrogen bond contributes twice to the analysis. Polar residues were assumed to be incapable of snorkelling.

Pairing of lipid-tail-accessible charges We tested the hypothesis that perhaps some charged residues prefer lipid-tail-accessible positions because they are paired and their charge is therefore somewhat neutralised. Within a hydrophobic environment, these paired charges would provide extremely strong bonds between the helices. This process does seem to occur, but it accounts for only 11% of the 1047 lipid-tail-accessible charged residues. It seems that for a charged residue to exist within a lipid-tail-accessible environment it need not necessarily be paired with an opposing charge if its hydrogen bonding requirements can be satisfied in other ways. 9% of lipid-tail-accessible charged residues form hydrogen bonds only with uncharged residues. This leaves 80% of lipid-tail-accessible charged residues that do not form hydrogen bonds with at least one other residue in the protein. These seemingly 'unsatisfied' residues are discussed later in this section.

Hydrogen bonding may occur either between two helices (interhelix) or between two residues separated by one turn on the same helix (intrahelix). The relative contributions of interhelix and intrahelix hydrogen bonding are shown in Figure 3.25 and an example of each is shown in Figure 3.26.

Interhelix hydrogen bonding is observed for 15% of lipid-tail-accessible charged residues. It may serve to anchor one helix relative to another, maintaining the correct conformation of the protein. Interhelical hydrogen bonds have been identified in the calcium ATPase with this role (Adamian & Liang, 2003). One example is shown in Figure 3.26(A). In this case, the paired charges occur between helices that otherwise share very few interacting residues, and probably function to anchor the helices together in the correct relative orientations. However, given the long and flexible side chains of arginine and lysine, this anchoring is unlikely to be highly rigid. An alternative role for interhelical hydrogen bonding may be to provide a driving force in the correct initial folding of a protein, particularly where the protein would otherwise have a tendency to mis-fold.

63 of the 1047 lipid-tail-accessible charged residues (6%) form intrahelix hydrogen bonds. An example is illustrated in Figure 3.26(B). Intrahelical salt bridges were first identified in globular proteins by Marqusee & Baldwin (1987), where they have been shown to increase the stability of helices. They also may serve to kink the helix, or to affect its flexibility, but this was not observed during visual examination of the dataset TM proteins. Chin & von Heijne (2000) have shown that charge interactions between lysine and aspartate residues placed one turn apart cause a polyleucine helix to be located further into the membrane. This suggests that the intrahelical pairing of the charges reduces the free energy change associated with membrane insertion, although the mechanism by which this occurs is unknown. Perhaps the interaction of charged and polar residues plays a



A - Interhelical B - Intrahelical

Figure 3.26: A: An interhelical ionic bond between the lipid-tail-accessible charged residues, His149C (red) and Asp36C (green) in ubiquinol oxidase (PDB code 1fft) (Abramson *et al.*, 2000). B: An intrahelical ionic bond between the lipid-tail-accessible charged residues, Asp59 (green) and Arg63 (red) in the calcium ATPase (PDB code 1eul) (Toyoshima *et al.*, 2000). This figure was produced using MolScript (©1997-1998, Per Kraulis).

similar role.

On average, each protein in the dataset has 12 inter-helical and 5 intra-helical hydrogen bonds (including those between both charged and polar residues), in addition to hydrogen bonds between main-chain atoms. Hence, almost one third of all hydrogen bonds (and a very similar proportion of all lipid-tail-accessible hydrogen bonds) formed between residue side-chains are intra-helical. Whilst the importance of inter-helix hydrogen bonding in TM proteins has recently been noted (Adamian & Liang, 2002, 2003), the significant contribution made by intra-helix interactions has not previously been recognised. This result has implications for the modelling of TM proteins. It suggests that, in general, the presence of charged or polar residues does not provide constraints by which helix-helix interactions can be predicted, since these residues need not be paired with other similar residues on adjacent helices.

Interaction of lipid-tail-accessible charges with water and head-groups As described above, approximately 80% of lipid-tail-accessible charged residues (842 out of 1047) do not form hydrogen bonds with any other residue in the protein. Of these, 25

(3%) formed hydrogen bonds with water. However, given the frequent difficulty in determining the precise location of water molecules in a structure, and the low resolution of many membrane protein structures, this figure may be affected by crystallographic errors.

The majority (532 or 63%) of the 842 unsatisfied residues were found close to the lipid-tail/head-group interface. It seems likely that these latter residues are able to reside favourably within the hydrophobic lipid-tail environment because they can 'snorkel' their charged groups up into the head-group region where they can form hydrogen bonds with water molecules or with polar atoms of the head-groups. An example is shown in Figure 3.27. Snorkelling can be clearly identified by visual inspection in many of the proteins analysed. The phenomenon is well known for both lysine and arginine (Killian & von Heijne, 2000; de Planque *et al.*, 1999; Mishra *et al.*, 1994), due to their long, positively charged side-chains. Further support for this hypothesis derives from the observation that, on average, 65% of the potential snorkelling residues in each protein are arginine or lysine. In contrast, a value of only 37% would be expected by chance alone, due to the relative frequencies of each residue type in the TM lipid-tail-spanning region as a whole. A possible role of snorkelling residues may be in vertically anchoring the protein in the membrane.

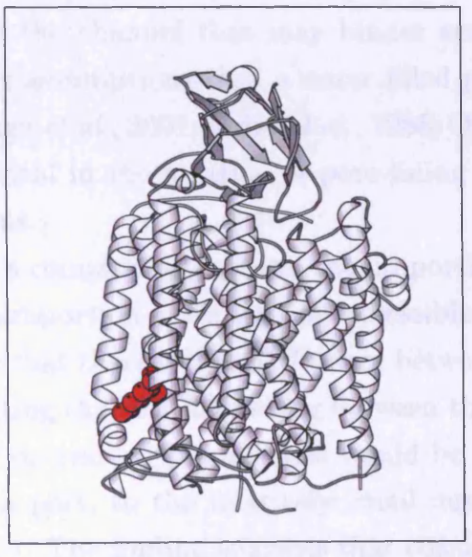


Figure 3.27: A 'snorkelling' lysine residue in cytochrome C oxidase (PDB code 1ehk) (Soulimane *et al.*, 2000). The Cα of lys16B (red) is located within the TM lipid-tail-spanning region but the charged side chain is able to reach the polar head-group region.

Role of lipid-tail-accessible charges in pore lining Charged residues are often required for the function of the protein, perhaps in the lining of an ion channel pore or the binding of cofactors or ligands. 10 of the 24 proteins in the analysis contain functional pores. Hence, it was postulated that some of the charged residues classified as lipid-tail-accessible, due to their large accessible surface area, may actually line a water-filled pore or cavity, and hence not be truly accessible to lipid-tails. However, only 75 of the 6835 lipid-tail-accessible residues (1%) were located within a channel. In addition, only 7 charged residues were found lining any of the pores, and these each made at least one interhelical hydrogen bond to another residue. Hence location within a water-filled pore is not solely responsible for satisfying the hydrogen bonding requirements of any accessible charged residues.

On analysis of the 75 pore-lining residues, it was found that only 9% were charged, 20% were polar, 7% were aromatic and 50% were non-aromatic hydrophobic residues. The most common pore-lining residues were isoleucine (16%), alanine (16%) and glycine (15%). The pore-lining residues are similar in hydrophobicity to buried residues (which also contain 50% of buried residues). The reason for the relatively high hydrophobicity of pore-lining residues is likely to be that they are suited to efficient flow of polar substrates, such as ions. This is because hydrophobic pore-lining surfaces prevent strong interactions between the substrate and the channel that may hinder transport. This finding is in contrast to several previous assumptions that a water-filled pore would tend to be lined with polar residues (Arechaga *et al.*, 2001; Milks *et al.*, 1988; Opella *et al.*, 1999; Oiki *et al.*, 1990). This knowledge is vital in the location of pore-lining helix faces when modelling pore-containing TM proteins.

Shown in Figure 3.28 is a comparison between the proportion of pore-lining residues of each residue type and the proportion of buried and accessible lipid-tail-spanning residues of each type. It can be seen that there is little difference between the profiles of pore-lining and buried residues, indicating that distinguishing between these groups of residues using a predictive method based on residue propensities would be very difficult. The problem is probably due, at least in part, to the relatively small number of pore-lining residues available in the dataset (75). The finding suggests that considerably more 3-dimensional structures of pore-containing TM proteins will be required before predictive methods can identify a channel's pore-lining residues from sequence alone.

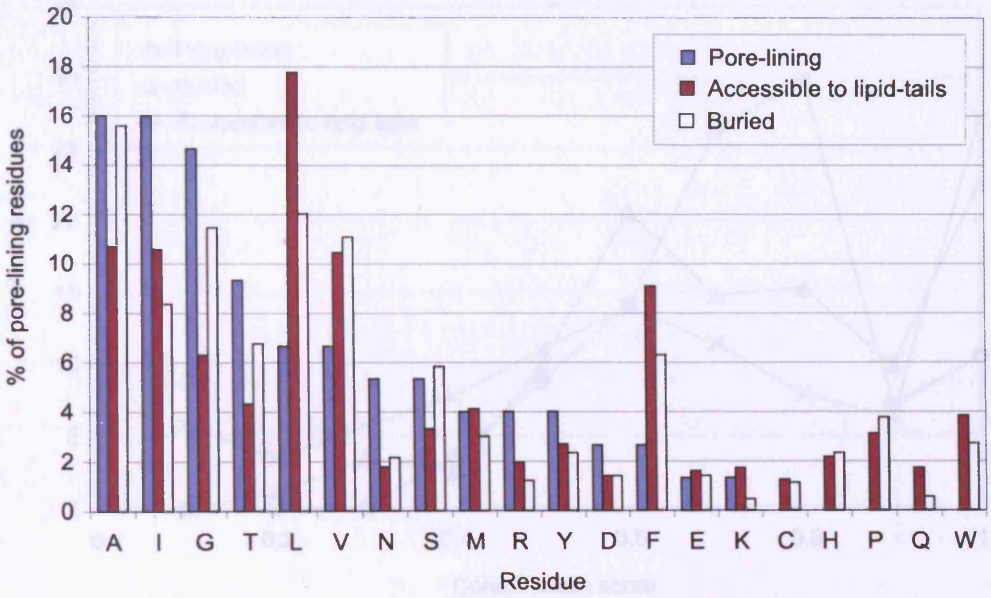


Figure 3.28: A comparison of the proportion of pore-lining, buried and lipid-tail-accessible residues of each residue type, in the lipid-tail-spanning region.

Similarly, there is no significant difference between the sequence conservation of pore-lining residues and buried residues (Figure 3.29, the average conservation score was 0.69 for pore-lining and 0.68 for buried residues). This result reflects the strong requirement for the conservation of both buried and pore-lining residues, albeit for different reasons. (Buried residues are likely to be conserved to maintain structure, while the conservation of pore-lining residues is important functionally). Hence, sequence conservation is unlikely to be of use in the identification of pore-lining residues. The fact that pore-lining residues are difficult to distinguish from buried residues by these methods may hinder the modelling of pore-containing TM proteins.

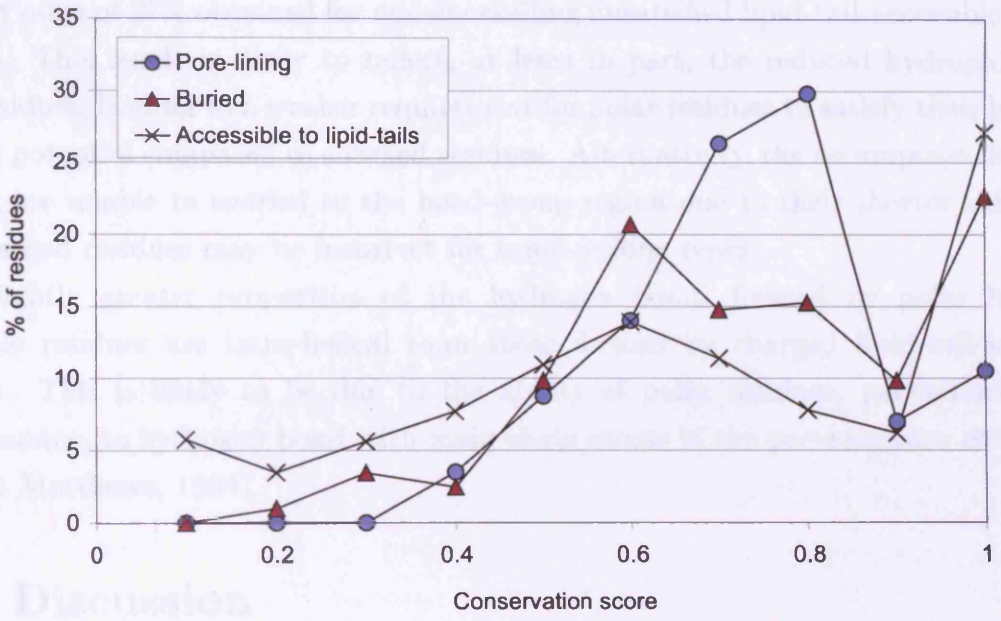


Figure 3.29: A comparison of the distribution of sequence conservation scores for 57 pore-lining residues, and residues in buried and lipid-tail-accessible positions in the lipid-tail-spanning region. The mean conservation is 0.68 for buried residues, 0.69 for pore-lining residues and 0.59 for lipid-tail-accessible residues. There is no significant difference between the conservation of pore-lining and buried or lipid-tail-accessible residues.

Non-hydrogen bonded lipid-tail-accessible charged residues The remaining 285 (27%) charged, lipid-tail-accessible residues appear to be genuinely accessible to lipid-tails. No source of hydrogen bonding has been identified for them by visual inspection of 3-dimensional structures, suggesting that they remain unsatisfied. It is possible that these residues form hydrogen bonds with cofactors or other molecules that are not found in the crystal structures. Alternatively, it may result from slight inaccuracies in the location of

the lipid-tail-spanning slice.

Hydrogen bonding patterns of lipid-tail-accessible polar residues

Figure 3.25 illustrates that, in general, polar lipid-tail-accessible residues form relatively similar hydrogen bonds to charged lipid-tail-accessible residues. Lipid-tail-accessible polar residues show a slightly greater tendency to form hydrogen bonds with polar and hydrophobic residues than do charged residues. It seems that the majority of hydrogen bonds made by charged and polar residues are to other residues of the same broad type.

77% of the polar residues lack hydrogen bonds to protein or water, considerably greater than the value of 27% obtained for non-snorkelling unsatisfied lipid-tail-accessible charged residues. This result is likely to reflect, at least in part, the reduced hydrophilicity of polar residues, leading to a weaker requirement for polar residues to satisfy their hydrogen bonding potential compared to charged residues. Alternatively, the assumption that polar residues are unable to snorkel to the head-group region due to their shorter side chains than charged residues may be incorrect for some residue types.

A slightly greater proportion of the hydrogen bonds formed by polar lipid-tail-accessible residues are intra-helical than those formed by charged lipid-tail-accessible residues. This is likely to be due to the ability of polar residues, particularly serine and threonine, to hydrogen bond with main-chain atoms in the previous turn of the helix (Gray & Matthews, 1984).

3.4 Discussion

During this chapter, the currently available polytopic TM protein structures were analysed and 24 TM protein families, represented in the protein structure databases, were identified. This number is more than twice that available for any previous analysis. Basic analysis of these structures were performed, generally confirming the results of previous, smaller studies by identifying the preferences of different residues for different TM environments. The results clearly show that the majority of TM helix-helix contacts are made by either small relatively polar residues, particularly glycine, alanine and serine, or large hydrophobic residues. The aromatic residues, and many of the charged and hydrophobic residues, show strong preferences for lipid-tail-accessibility.

This work confirms the results of Javadpour *et al.* (1999), using only 4 TM proteins, indicating that charged residues in the lipid-tail-spanning region show a preference for lipid-tail-accessible positions. However, the conclusions contrast with the more recent work of Ulmschneider & Sansom (2001), who stated a trend for the opposite preference.

Despite the larger dataset used for the latter analysis (15 TM proteins) the results are likely to have been biased by the inclusion of multiple members of the same protein family.

Possible explanations for the presence of lipid-tail-accessible charged residues have not previously been investigated. Here we show that the majority of lipid-tail-accessible charged residues are not paired, but do satisfy their hydrogen bonding potential in other ways, namely by snorkelling their charges into the head-group region, but also by interaction with other residue types. Interestingly, almost one third of the interactions with other residues are intrahelical.

As lipid-tail-accessible charged residues are present more often than would be expected by chance, it is likely that they confer some advantage to protein folding or function. For intrahelical paired charges, this may consist of increasing the stability of the protein, as has been shown for water-soluble proteins (Marqusee & Baldwin, 1987), and for a small polyleucine TM helix (Chin & von Heijne, 2000). The next steps will involve further experimental work to determine how this increase in stability is achieved and whether unpaired lipid-tail-accessible charged and polar residues have a similar role.

The thickness of the membrane varies between organisms, particularly between prokaryotes and eukaryotes, due to different lipid-tail compositions. For example, the hydrophobic length of a C22 phosphatidylcholine bilayer, as is optimal for the bacterial KcsA K+ channel, is thought to be approximately 34Å (Williamson *et al.*, 2003), whereas that for OmpF in a bacterial membrane is approximately 25Å (O'Keeffe *et al.*, 2000). These differences may have caused inaccuracies in the location of lipid-tail-spanning and head-group-spanning residues. At present it was felt that the dataset is not sufficiently large to permit division of the structures into prokaryotic and eukaryotic sets for separate analysis. However, this approach will likely prove an interesting study in the future, once more structures are available. The present study represents a set highly biased towards prokaryotes (78%), and it is therefore important to focus future structural genomics efforts more towards eukaryotic membrane proteins.

Many features of TM proteins were identified during this work that differ from those found in water-soluble proteins. TM helices are longer, more parallel and more highly conserved than the helices in water-soluble proteins. In addition, they contain different residues at buried and lipid-tail-accessible positions. The preferences of residues for buried or lipid-tail-accessible positions in TM proteins cannot simply be predicted by the use of a traditional hydrophobicity scale, since it is not true that all hydrophobic residues prefer lipid-tail-accessible positions and all hydrophilic residues prefer to be buried. A 'lipid-tail-accessibility scale' is developed during this work that represents the residue preferences found in TM proteins. Many poorly understood factors involved in residue positioning in TM proteins are encompassed in the LA scale, which shows very little correlation

with hydrophobicity scales. The LA scale, together with other knowledge gained during this work, will be of use in the prediction of TM protein structure in the future. The next chapter will investigate the use of these data for prediction of buried and lipid-tail-accessible residues, and use the results to formulate a model for the 3-dimensional structure of the uncoupling proteins.

Chapter 4

Modelling of uncoupling protein structure using hydrophobicity and conservation analysis of the primary sequence

4.1 Introduction

4.1.1 Aims

This chapter uses semi-automatic methods to identify the most likely arrangement of uncoupling protein transmembrane (TM) helices. It builds upon the results of both Chapters 2 and 3. Chapter 2 identified a list of possible TM helix (TMH) arrangements that are consistent with experimental and phylogenetic data for the family. It was concluded that no single arrangement of TM helices was more consistent with the currently available data than the others. Chapter 3 analysed the characteristics of TM proteins of known structure, and identified significant differences in hydrophobicity, residue type and evolutionary conservation between residues accessible to the membrane lipid-tails and those buried within the TM helix bundle. The importance of this finding lies in the fact that residue type, hydrophobicity and conservation are all sequence-based parameters that can therefore be obtained, and used in model prediction, for proteins for which no 3-dimensional structure is available.

This chapter attempts to use these observed differences between lipid-tail-accessible and buried residues to select between the models defined in Chapter 2 in a more rigorous manner. This is achieved by predicting, from their sequence characteristics, which residues

are likely to be buried and which are likely to be lipid-tail-accessible. An estimate of the accuracy of the method is made by performing the prediction on TM proteins of known structure. These predictions of residue location are then used to select the most likely model of helix arrangement from Chapter 2.

4.1.2 Motivation

The value of knowledge of the 3-dimensional arrangement of uncoupling protein TM helices can be divided into two main areas. Firstly, the ability to test the model via cross-linking experiments and other techniques will allow us to gain a greater understanding of the *structure* of the entire family. Secondly, the model will be of use in the guidance of mutational experiments, by identifying the likely functional residues, and this will lead to a greater understanding of the *mechanism* of UCP transport and regulation. On a broader level, this work has implications not only for increasing our understanding of the UCPs, but also of all other α-helix bundle TM proteins. Since this major class of proteins is thought to comprise 20-50% of most genomes (Arkin *et al.*, 1997; Wallin & von Heijne, 1998) and contains a huge number of potential drug targets, modelling techniques like this one will be of the utmost importance until high through-put TM protein structure determination is possible.

4.1.3 Previous work on TM protein structure prediction

Due to lack of structural data, TM structure prediction methods have often been based on sequence data (Donnelly *et al.*, 1993; Pilpel *et al.*, 1999) or on very small numbers of TM protein structures (Donnelly *et al.*, 1993). Some methods also analysed TM helices by comparing their characteristics to those of soluble proteins (Rees *et al.*, 1989). Taylor *et al.* (1994) used manual, helical wheel based methods in attempt to model the packing of bacteriorhodopsin TM helices. More recently, work in the area has concentrated on understanding both the mechanisms involved in TM helix interaction, by characterising the residues involved, (Rees *et al.*, 1989; Eilers *et al.*, 2000; Jiang & Vakser, 2000; Javadpour *et al.*, 1999; Ulmschneider & Sansom, 2001; Adamian & Liang, 2001) and the geometry of TM helices, in terms of length and angle (Bowie, 1997; Ulmschneider & Sansom, 2001). These studies have lead to considerable advances in the automatic prediction TM helix location, and, through the application of the Positive Inside Rule particularly, of TM protein topology (von Heijne & Gavel, 1988). It is now possible to identify TM helices and predict their topology with an accuracy of greater than 90% (Jayasinghe *et al.*, 2001; Jones *et al.*, 1994; Moller *et al.*, 2001).

Fleishman & Ben-Tal (2002) used knowledge of residue environment preferences to predict the likely arrangement of pairs of TM helices with a relatively high level of accuracy. (73% of predictions showed a root mean square deviation, compared to the actual structure, of less than 2Å). Pellegrini-Calace *et al.* (2003) have used position-specific membrane potentials to perform simulated annealing for the modelling of TM protein structure. However, due to high computational demands it is likely to be some time before the method is suitable for the modelling of larger TM proteins. Chen & Chen (2003) have used Monte Carlo folding methods to predict the seven-helix structure of rhodopsin with a secondary structure alignment error of 11%. However, each of these methods rely upon residue potentials and other data derived from soluble proteins. Given the considerable differences between the packing of TM and soluble protein helices (Rees *et al.*, 1989; Eilers *et al.*, 2000; Jiang & Vakser, 2000; Javadpour *et al.*, 1999; Ulmschneider & Sansom, 2001; Adamian & Liang, 2001), the use of data derived from soluble proteins is likely to limit the accuracy of these methods. Homology modelling of proteins belonging to a family in which at least one structure is known is another technique that has been used recently (Giorgetti & Carloni, 2003; Nikiforovich *et al.*, 2001). When this work was performed, however, no structure was yet available for any member of the mitochondrial carrier protein family, to which the UCPs belong.

More recently, structural data was used in the form of helix packing moments (Liu *et al.*, 2004c). However, the latter method was applied only to three protein families, and its effectiveness was not quantitatively assessed. In addition, the packing moments were not combined with any other form of data that may have improved the accuracy of the prediction considerably. The ways in which this predictive scale can be applied to generate and score TM helix arrangements in the modelling of TM proteins has also not been explored.

Two main attempts have been made at modelling members of the MCF. Nelson & Douglas (1993) used manual helical wheel variphobicity-based methods in combination with mutational analysis of the charged residues in the yeast ADP/ATP translocase. As described in Chapter 2, all members of the MCF show a pseudo-3-fold sequence repeat which indicates that they will show a pseudo-3-fold symmetric structure. While they adhered to pseudo-3-fold symmetry in the arrangement of the TM helices, Nelson & Douglas (1993) did not position homologous residues from each helix in equivalent positions. In addition, they modelled the family with one transport channel per monomer, despite the lack of evidence for or against this arrangement. More recently, Ledesma *et al.* (2002) produced a model for UCP1, using computational docking methods. However, the pair-potentials used were derived from soluble proteins, shown by many to differ considerably from TM proteins (Rees *et al.* (1989); Eilers *et al.* (2000); Jiang & Vakser (2000); Javad-

pour *et al.* (1999); Ulmschneider & Sansom (2001); Adamian & Liang (2001) and this work, Chapter 3). In addition, the model produced by Ledesma *et al.* (2002) lacked any form of 3-fold symmetry per monomer.

There is therefore a need for a simple, computationally inexpensive method for the prediction of TM protein structure for all families. Given the considerable differences between the packing of TM and soluble protein helices it should use structural data derived from TM proteins alone. Since the number of TM protein structures has doubled since the last comprehensive analysis (Ulmschneider & Sansom, 2001), considerable increases in accuracy should now be possible. In addition, a 3-dimensional model of the UCPs is required in which the functional unit conforms to the requirements that it exists as a dimer with pseudo-3-fold symmetry per monomer. Finally it will be important to compare the likelihood of models containing one and two transport channels to determine which is most likely to represent the native form of the protein.

4.1.4 Experimental evidence relevant to UCP structure

The possible models for UCP TM helix arrangements considered in this chapter are derived from the work of Chapter 2. Several assumptions have been made to limit the number of possible models. These assumptions are based on the experimental data outlined below, but are discussed more fully in Chapter 2, Section 2.2.1.

4.1.4.1 Evidence that the functional UCP is a dimer

As described in Chapter 2, evidence suggests that various members of the mitochondrial carrier family (Brandolin *et al.*, 1982; Bisaccia *et al.*, 1996; Kotaria *et al.*, 1999; Schroers *et al.*, 1998; Trezeguet *et al.*, 2000), including the UCPs (Lin *et al.*, 1980; Klingenberg & Appel, 1989), are dimers in their functional form. Much of this evidence is circumstantial, and most would not entirely rule out a higher-order oligomeric form. However, a higher oligomer is likely to be arranged as a dimer or trimer of dimers, with a functioning transport channel being formed by each dimer (Brandolin *et al.*, 1982; Huang *et al.*, 2001). The UCPs are hypothesised to function as dimers for the purpose of this work since constraints that limit the number of possible TM helix arrangements are badly needed. It is thought that, if a higher oligomer is correct, since the functional unit is thought to be a dimer, the model will remain valid.

In addition, several monomeric variations of the models are included for comparison. Since cross-linking and molecular weight determinations of native and functional expressed protein suggest a dimeric form, the interfaces between dimers are likely to be less tightly bound than the intra-dimer interfaces. If the dimer is in equilibrium with a higher order

oligomer, the inter-dimer interface is likely to have characteristics of both buried and lipid-tail-accessible residues, in order that it can be satisfied in both environments. This may help to minimise the problems associated with adopting a dimeric model if a higher-order oligomer proves to be correct.

4.1.4.2 Evidence that the functional UCP shows pseudo-3-fold symmetry per monomer

The requirement for pseudo-3-fold symmetry in the models derives from the tripartite domain structure of the UCP sequence. Each domain, comprising approximately 100 residues and including 2 TM helices (as described in Chapter 2, Section 2.1.6.1), is highly similar to the others. The tripartite structure is thought to derive from the triplication of an ancestral single domain protein. Since each domain is homologous, they are very likely to fold into similar structures, giving rise to pseudo-3-fold symmetry per monomeric unit.

4.1.4.3 Evidence that the functional UCP contains a single pore for transport

There is some evidence to suggest that the mitochondrial carrier proteins contain one transport pore per functional dimeric unit. This work has been carried out on the phosphate carrier (Schroers *et al.*, 1998) and the ADP/ATP translocase (Huang *et al.*, 2001). Due to the high sequence similarity between the members of the mitochondrial carrier protein family it seems valid to assume that, if some members of the family contain one pore per functional unit, the UCPs will also. However, since the evidence for a single pore is weak, a model containing two transport channels is included in the analysis for comparison with the single channel models.

4.1.4.4 Differences between lipid-tail-accessible and buried residues

Analysis of 24 TM protein structures in Chapter 3 lead to the conclusion that, in 64% of helices, residues buried within the TM helix bundle are both less hydrophobic and more conserved than those that are lipid-tail-accessible. While the level of conservation and hydrophobicity varied considerably among the helices the differences between buried and lipid-tail-accessible residues in each helix were highly significant. This illustrates that it is the difference in variphobicity between the buried and lipid-tail-accessible residues on a single helix that determines its folding, and not the absolute values. A cut-off based method would therefore not be suitable for identifying buried residues from conservation and hydrophobicity scores. A 'lipid-tail-accessibility' (LA) scale was also developed in Chapter 3, which was derived from the propensities of residues to be buried within the

helix bundle. This scale may also be used to distinguish buried from lipid-tail-accessible residues.

4.1.4.5 Geometric information from TM protein structure

The UCPs were assumed to contain a pore through which proton transport occurs, despite that fact that some proteins that transport protons, such as bacteriorhodopsin, do not contain an obvious pore. The assumption that the UCPs have a pore is based upon the knowledge that other members of the family, which are likely to show a very similar structure, transport large molecules like ATP, for which a pore would be required. In addition, Gonzalez-Barroso *et al.* (1997) have shown that if it's gating loops are removed, UCP function will a non-specific pore. Now that the structure of the adenine nucleotide carrier has been determined, this assumption has been validated.

The analysis of TM protein structure in Chapter 3 described the probable size of the pore and the number of TM helices likely to be lining it. A relationship was found between the total number of TM helices and the number of pore-lining helices, and between the number of pore-lining helices and the diameter of the pore. This information allows predictions to be made about the size of the pore and the number of pore-lining helices, purely from the knowledge that the UCP monomer contains 6 TM helices.

The work showed that, in a protein with 12 TM helices, as the UCP dimer is assumed to be, 6 of the TM helices are likely to contribute to the lining of the pore. In addition, the channel through such a protein is likely to be approximately 15Å in diameter (excluding side-chain volume). Alternatively, if a model is assumed in which each monomer forms a separate channel, 4 of the 6 UCP monomer TM helices are likely to contribute to the lining of the pore. This gives a pore diameter of approximately 10-15Å. These data provide support for Models 1 and 3t-u respectively, since these models have a helix arrangement that is most similar to TM proteins of comparable size and known 3-dimensional structure. Consistent with these predictions, the structure of the related adenine nucleotide carrier (Figure 4.18) has a funnel-shaped pore with an average diameter of approximately 15Å (calculated using the method described in Chapter 3, Section 3.2.4.6) and a maximum diameter of 20Å (Pebay-Peyroula *et al.*, 2003).

4.1.4.6 Proposed models for UCP TM helices

The models developed in Chapter 2 from the above information are illustrated in Figure 4.1. While the majority of evidence supports a dimeric, single pore structure, monomeric and two-pore models are included for comparison. It is hoped that knowledge of TM protein structure from Chapter 3 will permit these models to be ranked according to the

likelihood that they represent the actual UCP structure.

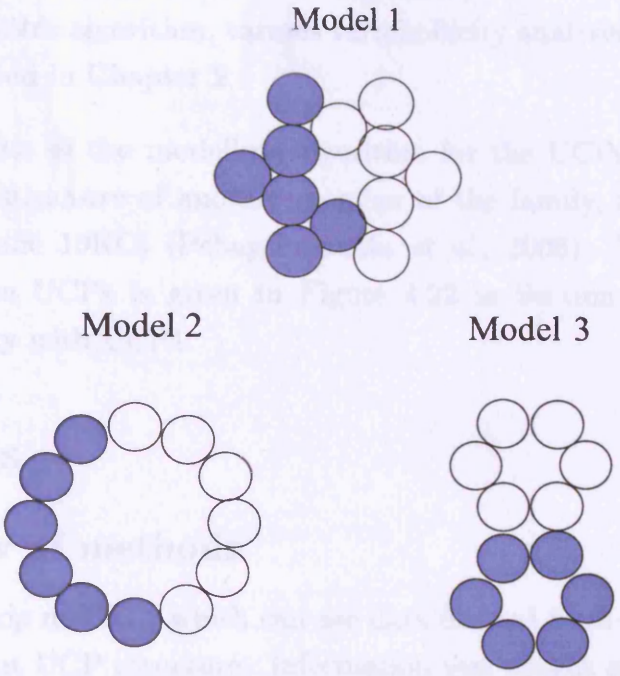


Figure 4.1: Models 1, 2 and 3: Possible arrangements of uncoupling protein transmembrane helices. Each helix, represented as a circle, has a diameter of 10Å. Each monomeric unit, one of which is shaded, consists of 6 TM helices. Model 1: 6 of the 12 dimer helices contribute to the pore, producing a pore with a diameter of approximately 10Å. Model 2: All of the 12 dimer helices contribute to the pore, producing a pore with a diameter of approximately 30Å. Model 3: Each monomer forms a separate transport channel with a diameter of approximately 10Å.

4.1.5 Summary

The aim of this chapter is to identify the most likely arrangement of the UCP TM helices from sequence alone. This will involve the following stages:

1. Development of a method for the prediction of buried and lipid-tail-accessible residues in TM helices, based on the principle that lipid-tail-accessible residues are less conserved, more hydrophobic and of different residue types than those buried in the helix bundle

2. Testing, evaluation and optimisation of this method using the dataset of known TM protein structures from Chapter 3

3. Use of this algorithm to predict which residues of UCP1 are more likely to be buried

4. To identify the most likely TM helix arrangement for the uncoupling proteins, based on the results of this algorithm, various variphobicity analyses and the experimental evidence described in Chapter 2

5. Lastly, the results of the modelling algorithm for the UCPs are compared to the recently solved structure of another member of the family, the adenine nucleotide carrier (PDB code 10KC) (Pebay-Peyroula *et al.*, 2003). The alignment of this protein with the UCPs is given in Figure 4.22 in Section 4.4.1. It shows 20% sequence identity with UCP1.

4.2 Methods

4.2.1 Overview of methods

The goal was to develop methods which can use data derived from sequence alone to provide information about UCP structure. Information was sought about the location and role of both particular residues and whole helices, in order that one model could be identified as the most likely. The first task was to identify the buried and lipid-tail-accessible faces of TM helices, using various characteristics of the residues, including hydrophobicity and sequence conservation. This enabled the helices to be orientated within each potential model. The second stage was to discriminate between the proposed models, by providing an overall 'score' for each potential structure, according to its compatibility with the sequence data. In addition, the probable location of each TM helix (pore-lining, buried within the protein or peripheral to the helix bundle) was identified from its sequence characteristics and combined with experimental data to provide evidence for or against each model. This strategy is summarised in Figure 4.2.

Analysis of TM protein structures has shown that residues buried within the TM helix bundle are smaller, (Jiang & Vakser, 2000) less hydrophobic and more conserved than lipid-tail-accessible residues, and they also contain a different distribution of residue types (Chapter 3). Major differences between the rules governing the location of particular residue types in TM and soluble proteins have been identified. A 'lipid-tail-accessibility scale' was also derived in Chapter 3, reflecting the propensity of each residue to be buried in the lipid-tail-spanning region of TM proteins. This knowledge was used in an attempt to distinguish buried from lipid-tail-accessible residues in the UCPs and select between the proposed models. The method was developed and tested using 23 of the 24 TM

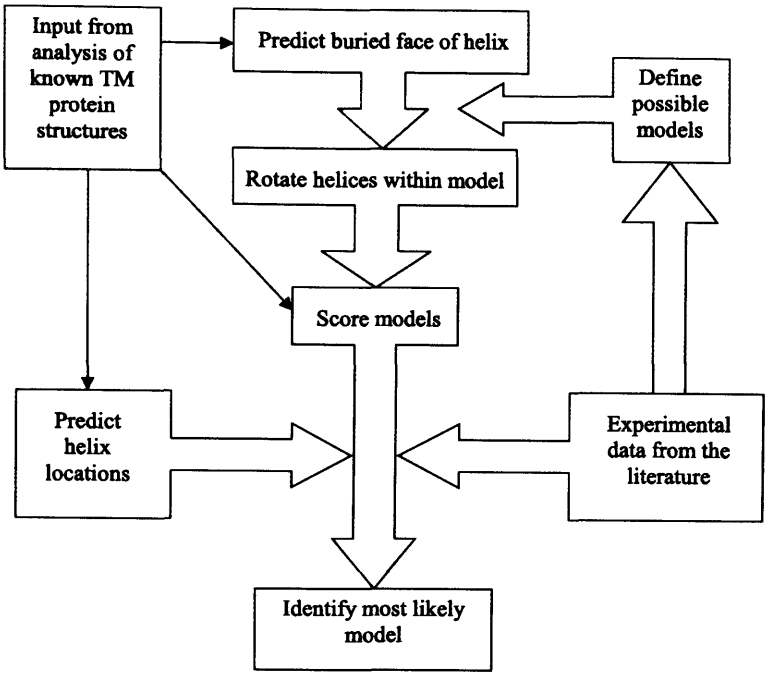


Figure 4.2: Flow diagram showing the strategy used to model the UCPs.

protein structures described in Chapter 3, Section 3.3.1. (The structure of the adenine nucleotide carrier was excluded since it belongs to the same family as the UCPs).

The potential models were scored according to the degree of correlation between the position of each residue in the model and its characteristics of conservation, residue type and hydrophobicity. The methods used in this chapter therefore can be divided into 4 stages, which are described in turn in this section:

1. Development of the prediction algorithm
2. Assessing the accuracy of the algorithm
3. Use of the algorithm to identify buried residues
4. Model generation and scoring

4.2.2 An algorithm to predict the buried face of TM helices

The algorithm used for the prediction of buried residues from sequence considers each helix as a separate unit. Each helix is represented as a helical wheel, as first described by Schiffer & Edmundson (1967), with centre at point (0,0). Each residue around the

circumference of the helical wheel is represented as a vector, with magnitude equal to the parameter used for scoring. The direction of the vector is determined by the position of the residue on the helical circumference. A vector sum is used to combine the individual vectors for each residue into a single helix vector which indicates the predicted buried face of the helix. The method is derived from the calculation of hydrophobic moments used by Eisenberg *et al.* (1982b).

Various parameters, described in Section 4.2.2.1, were used for scoring including the LA scale developed in Chapter 3, the Kyte and Doolittle (Kyte & Doolittle, 1982) (KD) and White and Wimley (Wimley *et al.*, 1996; Jayasinghe *et al.*, 2001) (WW) hydrophobicity scales, residue volume (Tsai *et al.*, 1999) and sequence conservation. In the case of conservation and residue volume, the values assigned to each residue type were scaled to lie between 0 and 1. Highly conserved or small residues therefore received a score close to 1, indicating that they are likely to be buried. These residues would make a large contribution to the size and direction of the helix vector. In contrast, poorly conserved or large residues received a score close to 0, and would be able to make little contribution to the helix vector in that direction. The KD, WW and LA scales were, however, scaled to lie between -0.5 to +0.5. This allowed hydrophobic residues, or those with a tendency to be lipid-tail-accessible, to make a negative contribution to the helix vector in that direction.

The sequence positions of the UCP TM helices were taken from Aquila *et al.* (1985), who predicted helix location using hydrophobicity. Only the central 18 residues of each helix were used in order to minimise the effect of inaccuracies of TM helix location, which may have lead to the inclusion of head-group-spanning residues. A variation of the prediction algorithm was developed for the UCPs, which allowed a single prediction to be made for each of the two groups of homologous helices. This was achieved by basing the prediction upon the mean parameter values of the three helices in each group. The parameters used for the UCP buried helix face prediction were those identified as giving the most accurate prediction for the proteins of known structure.

4.2.2.1 Scales used for prediction

Three different scales were used for prediction, as described below.

1. The White and Wimley (WW) scale (Wimley *et al.*, 1996; Jayasinghe *et al.*, 2001) is calculated from the differential partitioning of residues between water and n-octanol, and therefore represents purely a chemical property, the hydrophobicity, of the residues. Only if hydrophobicity is the major driving factor in TM helix packing would this scale would be expected to perform well in prediction.

2. The Kyte and Doolittle (KD) scale (Kyte & Doolittle, 1982) is derived from a combination of residue accessibility scores in soluble proteins and water-vapour transfer free energies. Hence the KD scale contains information about the environment preference of the residue, in addition to its hydrophobicity. However, the relevance of this information concerning the preferred accessibility of residues in soluble proteins to prediction in TM proteins may be limited, considering the known differences between these two classes of protein.

3. A lipid-tail-accessibility (LA) scale was derived from the TM-lipid-accessible/buried propensities observed in the analysis of TM protein structure described in Chapter 3. The value that a residue receives in the empirical scale is computed by proportionally scaling the residue propensities into the desired range of -0.5 to +0.5. Hence, this is not a traditional hydrophobicity scale, representing simply the water/lipid solubility of a residue, but a measure that encapsulates all of the factors affecting the positioning of a residue in a real TM helix bundle. It therefore seems likely that this scale will give the strongest predictions. As with all of the parameters used for prediction, the higher its value the greater the tendency for a residue to be found in buried positions.

4.2.2.2 Conservation scoring methods used

All conservation scores were computed by SCORECONS (Valdar & Thornton, 2001), using all default parameters, as described in Chapter 1, Section 1.3.2. For the UCPs, analyses were carried out using conservation scores derived from the whole mitochondrial carrier protein family, the UCP subfamily and the ratio of UCP/MCF conservation. MCF-derived conservation scores were generated using the 170 non-redundant sequences retrieved from a PSI-BLAST search, using a threshold of $1x10^{-40}$. This relatively high cut-off was used to ensure that function was preserved amongst all homologues identified. UCP-derived scores were calculated for 30 of these sequences, known to belong to the UCP subfamily. For the TM proteins of known structure, analyses were carried out at one level only, also using a threshold of $1x10^{-40}$.

4.2.3 Assessing the accuracy of the prediction

It is desirable to assess the performance of the prediction algorithm for two reasons. Firstly, it gives an estimate of the confidence with which predictions for the UCPs can be made. Secondly, it enables different prediction methods to be compared, particularly various methods of conservation and hydrophobicity scoring, in order to optimise the

performance of the algorithm.

The performance of the algorithm was assessed, using a jack-knifing procedure, by performing the prediction on the 23 of the 24 TM proteins described in Chapter 3, Section 3.3.1 (the adenine nucleotide carrier was excluded). Since the structures of these 23 proteins are known, the relative accessible surface area (%ASA) calculated by NACCESS v2.1.1 (©, S. Hubbard and J. Thornton, 1992-1996) was used to identify the correct buried and accessible faces of the helix. This is achieved in a similar way to that described above, in which a vector sum is taken of the residue vectors, the magnitude of each is determined by their 1-%ASA. (1-%ASA is used, not %ASA, in order to identify the most buried, not the most accessible, helix face).

Performance was assessed by calculating the average the difference in angle between the helix vectors generated from the predictive parameter and from the %ASA. This value is referred to as the angular error. Since a jack-knifing procedure was used, the data derived from each protein were excluded from the set used to predict the buried helix faces of that particular protein. This method maximises the amount of data available, to increase the accuracy of the prediction, while allowing the prediction parameters to be independent of the protein used for testing.

The various parameters used for prediction, described in the previous section, were used alone and in combination to optimise the performance of the algorithm, by minimising the mean angular deviation over all 178 helices in the dataset. Predictions were made and tested for the TM proteins of known structure using not only the whole helix, but also the centre of the helix with 1, 2, 3 or 4 residues removed from either end. This allowed the most effective method to be identified for the use in the UCP prediction.

4.2.4 Using the algorithm to identify the most buried face of the UCP TM helices

Multiple sequence alignments, annotated with conservation scores by SCORECONS, were generated for the entire mitochondrial carrier protein family, using the methods described in Section 4.2.2.2. This data was then used as input for the prediction algorithm developed in this chapter, to identify the likely buried face for each of the UCP TM helices. The predictions were based on family-derived conservation, KD hydrophobicity and LA scale scores and the terminal 2 residues from each end of the helices were excluded. This method was adopted since it gave the best performance when tested on the 23 TM proteins of known structure, as discussed in Results Section 4.3.2.

4.2.5 Model generation and scoring

Each of the TM helix models described in Chapter 2 were scored according to their compatibility with different types of data. This enabled them to be ranked in order of the likelihood that they represent the actual UCP structure. The scoring methods are described in the following sections.

1. *Scoring against experimental data*

   The most likely model can be selected by assessing the degree of correspondence between the residue locations of the model and the experimental evidence for their position and role shown in Table 4.1. Models were generated by rotating each helix until the buried face identified by the algorithm was buried within the model helix bundle and the opposing face was accessible to membrane lipid-tails. Since no scoring system can be employed with so few pieces of data, the models are simply ranked according to the number of pieces of experimental evidence with which they are consistent.

Mutant	H^+	Cl^-	NTP	pH	Notes
D27N	+	-	-	-	Probably not directly involved in transport, instead required to maintain native conformation (Echtay *et al.*, 2000a; Urbankova *et al.*, 2003)
D27E	-	-	-	-	Shows need for negative charge (Echtay *et al.*, 2000a; Urbankova *et al.*, 2003)
H214N/W	-	-	-	+	When unprotonated blocks NTP binding (Echtay *et al.*, 1998)
D209N	-	-	-	+	Retract H214 when it is protonated to allow NTP binding (Echtay *et al.*, 2000a)
D210N	+	-	-	+	Retract H214 when it is protonated to allow NTP binding (Echtay *et al.*, 2000a)
R83I R182Q R267Q	-	-	+	-	Suggested to bind nucleotide phosphate, but essential for whole MCF suggesting not a direct role Echtay *et al.* (2001a); Modriansky *et al.* (1997)
R91T	-	+	-	+	Thought to influence nucleotide binding via an H bond to E190, but this H bond is impossible (Echtay *et al.*, 2001a)
Cysteines	-	-	-	-	None of 7 cysteines required for activity (Arechaga *et al.*, 1993).
C24 + D27 + T30					Triple mutant fully non functional (Arechaga *et al.*, 1993).
W280					Quenching suggests location in water-filled cavity. May bind nucleotide purine ring in pore (Jezek *et al.*, 1998).

Table 4.1: Summary of the results of a range of mutagenesis experiments, describing the mutation, its effect on UCP function and its consequent implications for the location of the residue concerned. Column headings 'H^+', 'Cl^-', 'NTP' and 'pH' refer to H^+ and Cl^- transport, nucleotide binding and the pH sensitivity of nucleotide binding respectively. A '+' indicates that this factor is affected by the mutation, whereas a '-' indicates that the factor shows a wildtype phenotype despite the mutation.

2. *Scoring against conservation and hydrophobicity data and LA score*

The models presented in Figure 4.1 are very schematic, representing only helical topology. A key step is to 'translate' these models into residue-based models, in which each residue can be assigned as either lipid-tail-accessible or buried. To achieve this the 'buried positions' were defined for each model, using the method shown in Figure 4.3. Lipid-tail-accessible and pore-lining residues were those that remained, adjacent to the membrane lipid-tails or pore respectively, once all buried residues had been identified. The pore-lining residues were then grouped with the buried residues, due to their similar characteristics of conservation and residue propensity, as described in Chapter 3. However, improvements in accuracy may be possible when more buried residues have been analysed and their distinct characteristics more clearly defined.

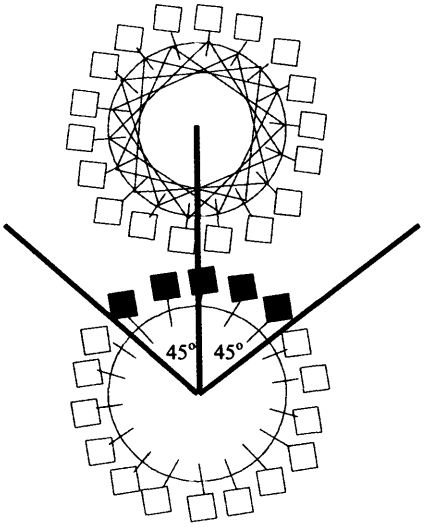


Figure 4.3: Definition of buried positions within each model. The residues buried between two helices are defined to be those found within a 90° arc (red lines) centred on the line connecting the centres of those two helices (black line).

Next, each residue type was assigned two scores per parameter, representing the probability that a residue of that type is found in a buried or accessible position, given the value it obtains according to that parameter. For example, if scoring based on the hydrophobicity parameter, a hydrophobic residue such as leucine would obtain a high score if lipid-tail-accessible (a favourable position), but a low score if buried (an unfavourable position). These are referred to as 'residue type scores'. For the conservation parameter, rather than residue type scores, each sequence position received a different score according to its particular conservation, as described below.

The scores of all of the residues in a model are summed, taking the relevant value, depending upon whether that residue is buried or accessible in that model. Since a high residue score indicates that the residue is found in the most likely environment, the sum of all the residue scores generates a 'model score' representing the degree of fit shown between the model and the data. This value can be used to assess the likelihood that that model is the correct one.

The parameters used for prediction were KD hydrophobicity, LA score, residue volume and sequence conservation, both alone and in combination. For residue volume and KD hydrophobicity, the residue type scores were simply the parameter values, scaled to run from -0.5 to 0.5, for buried residues and the same values multiplied by -1 for accessible residues. For the LA scale the propensities to be buried or lipid-tail-accessible, as described above, were used. For residues in buried positions in the model, the conservation-based score was taken as the conservation score calculated by SCORECONS, subtracting 0.5 to give equal weighting with the other parameters used. The value for accessible positions was taken as the SCORECONS score, subtracting 0.5 and multiplied by -1, representing the idea that lipid-tail-accessible residues have a low probability of being highly conserved.

For each model, each helix was rotated in turn and a score calculated for all of the resulting possible arrangements. Homologous helices were constrained to be located in equivalent rotational positions, in order to maintain pseudo-3-fold symmetry. Model scores were then generated by summing the scores of the individual helices. The score quoted for each model is that obtained for the highest scoring helix arrangement with 3-fold symmetry. In the case of the score based upon all parameters combined, the models were scored using the average of the residue scores from all parameters.

Theoretically, the maximum score using this method is 54 and the minimum possible is -54. However, these values could only be obtained if all 108 residues in the model obtained the maximum or minimum score of 0.5. These values are clearly unobtainable, particularly for the combined score, since the different scoring parameters have different residue types with maximum and minimum scores.

3. *Scoring against UCP/MCF conservation ratio data*

Residues with a high UCP/MCF conservation ratio (Section 4.2.2.2) are likely to play functionally important roles specifically in the UCPs, and hence will be enriched in pore-lining residues. In contrast, residues showing high conservation across all sequences are likely to play general structural roles throughout the whole MCF

family. Hence these data can also be used to score the likelihood of the UCP models.

The 20 residues with highest UCP/MCF conservation ratio and the 10 residues with highest conservation in the MCF were used for scoring. The buried, lipid-tail-accessible and pore-lining positions for each model were defined as described previously. Predictions were considered correct for (1) residues with high UCP/MCF conservation ratios (functional), found in either a buried or pore-lining positions and (2) residues conserved throughout the family (structural) that are found in buried (but not pore-lining) positions. A score was then calculated for each model, as:

$$\text{Score}_X = \frac{\text{N}_F + \text{N}_S}{\text{N}} \tag{4.1}$$

Where:

Score_X = Score for Model X.

N_F = Number of correctly predicted high UCP/MCF conservation ratio residues (functional).

N_S = Number of correctly predicted high MCF conservation score residues (structural).

N = Total number of residues analysed = 30.

Hence, a score of 100 would represent a perfect fit between the model and the data, with every residue found in the environment expected from its conservation ratio.

4.2.6 Determining the position of TM helices in the helix bundle

Information concerning the location of the UCP TM helices would be of use in selecting between the proposed models or validating the selection made by other methods. It was therefore necessary to investigate the ability of various parameters to predict the location, either buried in the helix bundle or accessible to lipid-tails at the periphery, of the 138 helices in the TM proteins of known structure. The average %ASA was calculated for each helix by taking the mean of the %ASAs for each residue in that helix.

The distribution of mean helix %ASA values was approximately normal with a mean of 22%, so that there was no clear cut-off which would allow buried and lipid-tail-accessible helices to be easily distinguished. A cut-off of 12% ASA was therefore arbitrarily selected and used to determine the proportion of buried helices for each protein. On average, 30% of helices in each protein had a mean %ASA of less than 12% and were classified as buried, while the remaining 70% of helices were classified as lipid-tail-accessible. If each protein is

omitted in turn, in an attempt to jack-knife the results, the average proportion of buried helices varied from 27%, for all proteins except 1um3, to 33%, for the set excluding 1jb0. Given the small number of TM helices found in each protein, these variations had no effect on the number of helices classified as buried, indicating that no particlar protein biases the dataset. The mean value of 30% was therefore used for all proteins.

Next, the ability of various parameters to predict helical location was tested. The average conservation, residue volume, KD, WW and LA scale score were calculated for each helix, by taking the mean of the values for each residue. The same proportion of buried and lipid-tail-accessible helices was predicted as had resulted from the classification based on %ASA (30%). Taking conservation as an example, 30% of the helices with the highest conservation score in the protein were predicted to be buried, and the remaining helices were predicted to be lipid-tail-accessible. The process was repeated for all of the 23 proteins in the dataset.

If a helix was both predicted by its conservation score to be buried, and classified as buried due to mean %ASA of less than 12%, it was considered a true positive. Similarly, if a helix was predicted by both its conservation score and its mean %ASA to be lipid-tail-accessible, it was considered a true negative. Conversely, an incorrectly predicted helix was considered either a false positive or negative. The effectiveness of the method was assessed by calculating a % accuracy as the number of true positives plus true negatives, divided by the total number of helices. The method was then repeated to test the ability of residue volume, KD, WW and LA scale score to predict helix location. Combined scores were obtained by summing the helix scores according to each individual parameter. The results of the analysis are given in Table 4.2 in the Results section.

The next stage was then to use this information to make a prediction about the location of TM helices in the UCPs. Since a combination of all parameters (conservation, LA, WW and KD scale and residue volume), termed the 'all-parameter combined score', demonstrated the greatest predictive ability in proteins of known structure, this method was used. The 2 UCP helices (30% of 6 TM helices) with the highest score were predicted to be buried.

4.3 Results

4.3.1 Overview of results

The results for this work can be divided into five main parts. The first is the analysis of the accuracy of the prediction algorithm developed. This is followed by the results of the optimised algorithm in the form of helical wheels of each UCP TM helix indicating

their predicted buried faces as vectors. Various variphobicity analyses are then described, including hydropathy profiling of whole TM helices and a comparison of residues that are conserved across the whole MCF and the UCPs alone. A discussion of the likelihood of each of the models described in Chapter 2 is given, based on the predictions from the algorithm, the detailed variphobicity analysis and experimental mutagenesis work. Finally, the proposed model is compared to the actual structure of a related protein that has recently been solved, the adenine nucleotide carrier.

4.3.2 Effectiveness of the prediction algorithm at identifying the most buried helix face

Figure 4.4 gives a summary of the performance of the algorithm using various combinations of parameters. Removing one residue from either end of each helix increased the average performance for all parameters, as measured by the mean angular error, from 23° to 19°. Using this method, the combination of LA scale and conservation gave the greatest accuracy, with an average angular error between predicted and actual buried vectors of 11°. However, this method was greatly affected by removal of residues from the helix ends, suggesting that it may be very sensitive to the accuracy of locating the helices in the sequence. Consequently, KD hydrophobicity, LA scale and conservation has been selected for use for performing the UCP prediction, since this method gave very high accuracy (13°) and very similar results for different helix lengths. As an illustration of the high accuracy of the method, 95% of the 178 TM helices in the dataset were predicted to within 25° of the actual vector. Furthermore, the performance of this method is expected to increase, as more TM protein structures are solved, and the standard errors associated with the LA scale are reduced.

The accuracy of even the best similar scales for prediction of buried TM helix faces is considerably less than that for our LA scale. For example, the mean angular error for helix centres (one residue removed) is 16° for the LA scale alone, 27° for the helix packing scale developed by Liu *et al.* (2004c) and 46° for the Kprot scale (Pilpel *et al.*, 1999).

The LA scale alone gives a very similar accuracy to the KD. This is unexpected, since the LA scale is derived from an analysis of TM residues, where as the KD scale is based on accessibilities in soluble proteins and on water/vapour partition coefficients. The reasons for this seem to be that those residues with very high or low scores on the LA scale, and therefore that have the most discriminatory power during prediction, are very infrequent. Most of the very common residues have scores close to 0 on the LA scale, and will therefore be able to contribute little to the accuracy of the prediction. In contrast, residues with high discriminatory power on the KD scale, like the very hydrophobic residues I, L and

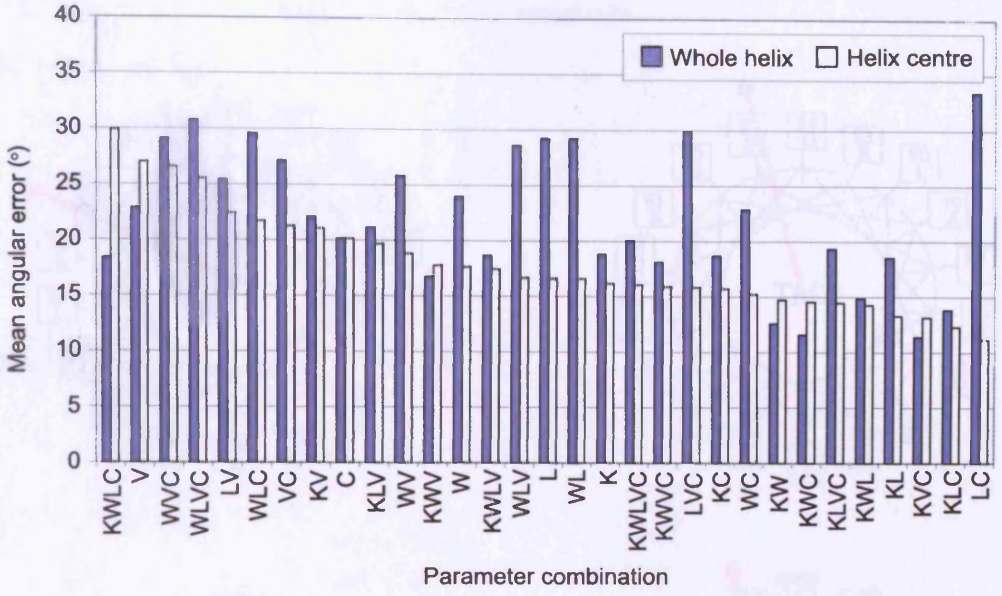


Figure 4.4: Average angular error scores for 178 non-homologous TM helices of known structure using various combinations of parameters for prediction. See Section 4.2.3 for a description of the calculation of angular error. 'Helix centre' indicates that one residue has been removed from either end of the helix. C = sequence conservation; W = White and Wimley hydrophobicity; V = residue volume; L = lipid-tail-accessibility scale; K = Kyte and Doolittle hydrophobicity.

V, are extremely common. This may explain the unexpectedly high performance of the KD scale relative to the LA scale.

Despite this, it is apparent that the LA scale carries some different and complementary information to that in the KD scale. For example, the average angular error for the helix centre is 16° for KD scale but 13° for the KD and LA scales combined. This indicates that the LA scale will make a valuable contribution to TM protein modelling, particularly as more structures become available and the accuracy with which the LA scale is defined increases.

4.3.3 Predicted TM buried residues of UCP1

The buried face of the UCP TM helices, as identified by the prediction algorithm, are shown in Figure 4.5. Crucially, it can be seen that homologous residues are found on the buried face of TMHs 1, 3 and 5 and TMHs 2, 4 and 6. This is important because it implies that the homologous helices do occupy equivalent positions and therefore that the real structure does indeed show pseudo-3-fold symmetry.

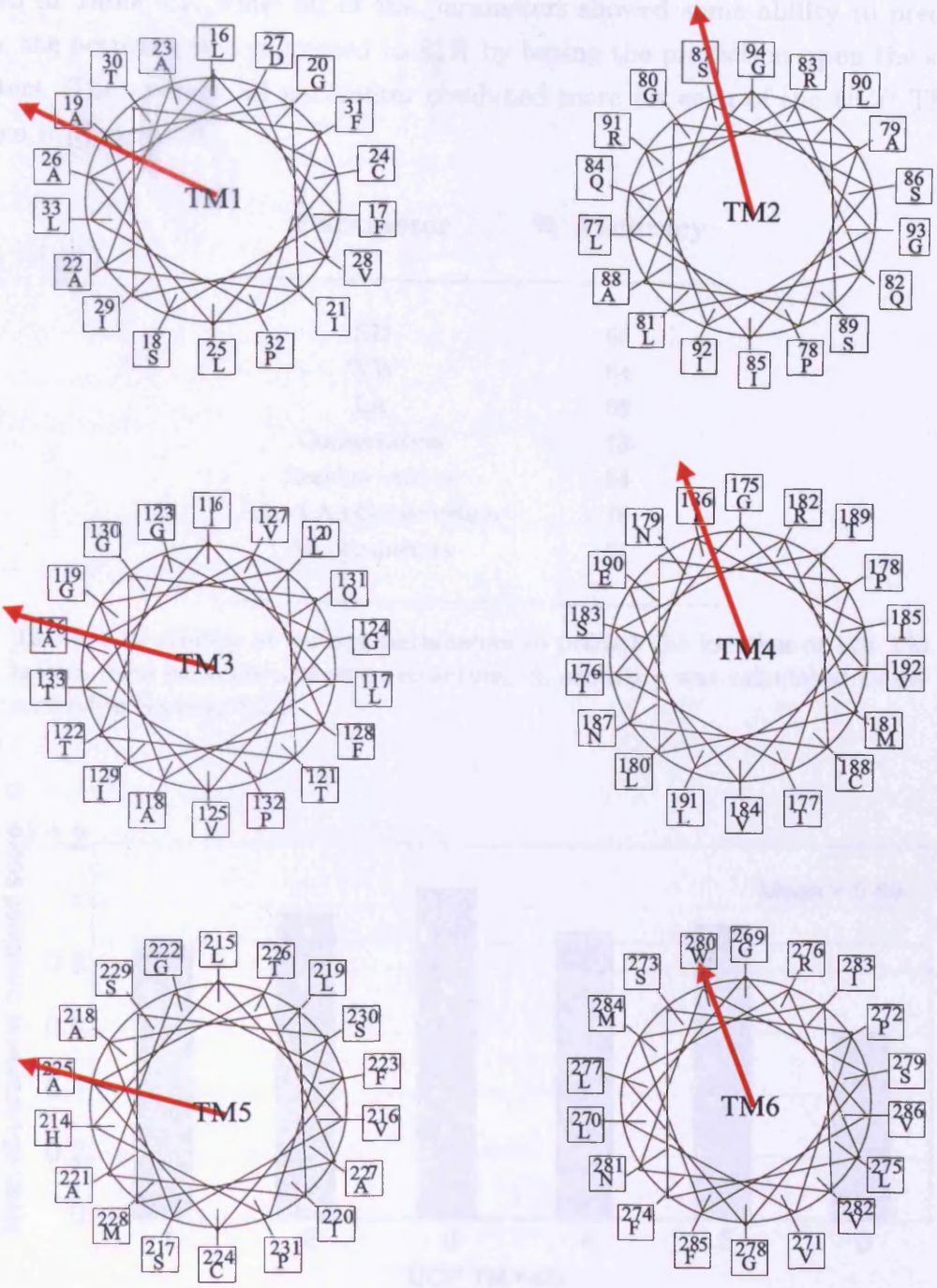


Figure 4.5: Helical wheels of the UCP TM helices with arrows (red) showing the predicted buried face. The helices are aligned by sequence to illustrate the homologous position of the buried vector in helices 1, 3 and 5 and in helices 2, 4 and 6.

4.3.4 Predicting the likely position of UCP TM helices

As shown in Table 4.2, while all of the parameters showed some ability to predict helix location, the accuracy was increased to 81% by basing the prediction upon the sum of all parameters. The average all parameter combined score for each of the UCP TM helices are shown in Figure 4.6.

Parameter	% Accuracy
KD	68
WW	64
LA	66
Conservation	73
Residue volume	64
KD+LA+Conservation	78
All parameters	81

Table 4.2: Ability of various parameters to predict the location of 178 TM helices from proteins of known structure. % accuracy was calculated as described in Section 4.2.6.

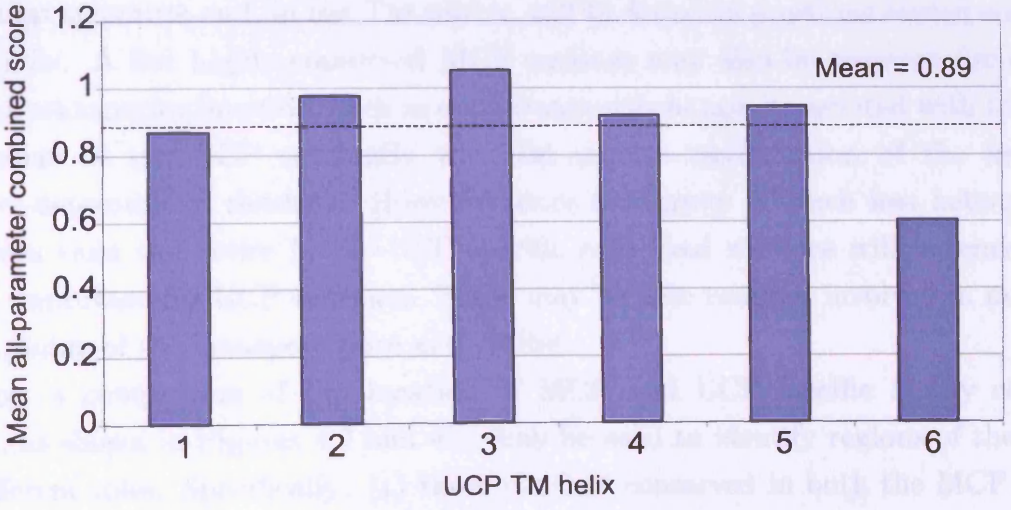


Figure 4.6: Sum of average all-parameter combined score for each UCP TM helix. The mean of 6 individual helix values is indicated by the dashed line.

The methods used successfully to predict helix location for proteins of known structure were based on the assumption that approximately 30% of the TM helices in each protein

tend to be buried within the helix bundle, equivalent to 2 of the 6 UCP helices. As shown in Figure 4.6, TM helix 2 and, particularly, 3 are the UCP helices with the highest combined scores, and hence are the most likely to be buried. There is no obvious pattern in the scores received for each helix and the even- and odd-numbered helices do not appear to differ in their scores. Therefore, in order to maintain pseudo-3-fold symmetry, evidence suggests that all helices are located in equivalent positions. If this were the case, rather than being due to differences in helix location, any variation in the score of the helices must either be explained by random noise or by other, perhaps functional, constraints. One possibility is that the helices form a ring, but that the monomer-monomer interface is formed mainly by helix 3, decreasing its accessibility to lipid-tails relative to the others. Conversely, the lower score of TM 6 could be explained by its location on the opposite face of the monomer to the interface, leading to its relative protrusion into the lipid-tails.

4.3.5 Information derived from family- and subfamily-specific conservation scores

4.3.5.1 Overview

Since all members of the MCF are of common evolutionary origin, they will share a common 3-dimensional structural fold, despite their divergent functions. Hence, residues highly conserved throughout the whole family are likely to play important roles in maintaining that structure and, in the TM region, will be found in positions buried within the helix bundle. A few highly conserved MCF residues may also be required for common aspects of transporter function, such as conformational changes associated with transport.

Members of the UCP subfamily will also require conservation of the important 'structure-determining' residues. However, since this group is much less heterogeneous in function than the entire MCF, UCP-specific conserved residues will be enriched for residues important for UCP function. These may include residues involved in nucleotide binding, lining of the transport pore and gating.

Hence, a comparison of the location of MCF and UCP-specific highly conserved residues, as shown in Figures 4.7 and 4.8, may be used to identify regions of the protein with different roles. Specifically: (1) those residues conserved in both the MCF and the UCPs are likely to play a general structural role, (2) those conserved in the UCPs alone, identified by a high UCP/MCF conservation ratio, are more likely to have a UCP-specific functional role.

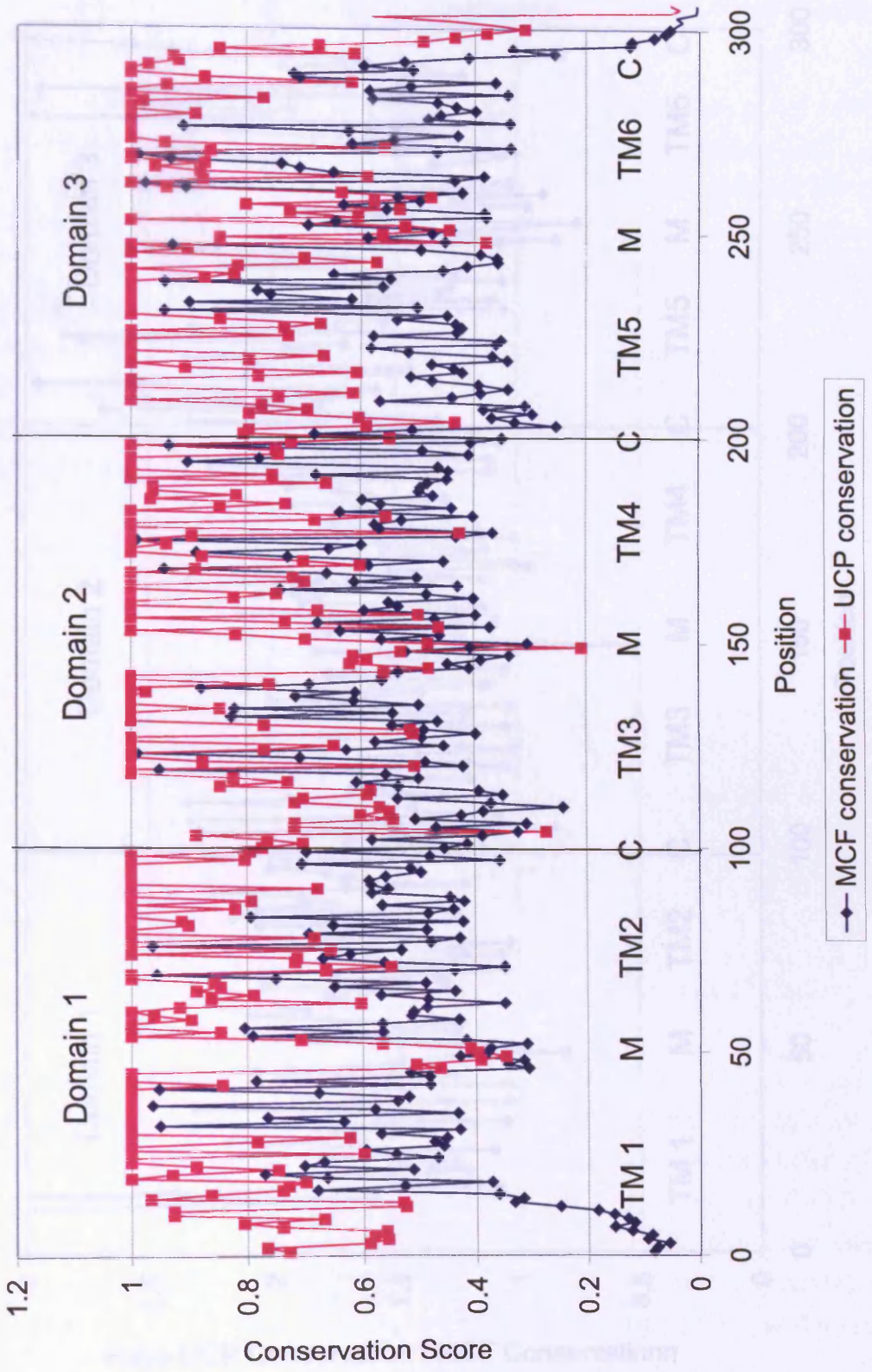


Figure 4.7: Comparison of UCP-derived conservation score and MCF-derived conservation score along the length of the UCP sequence. Horizontal lines divide the UCP sequence into its 3 homologous domains. TM : transmembrane helix, C : cytosolic loop, M : matrical loop.

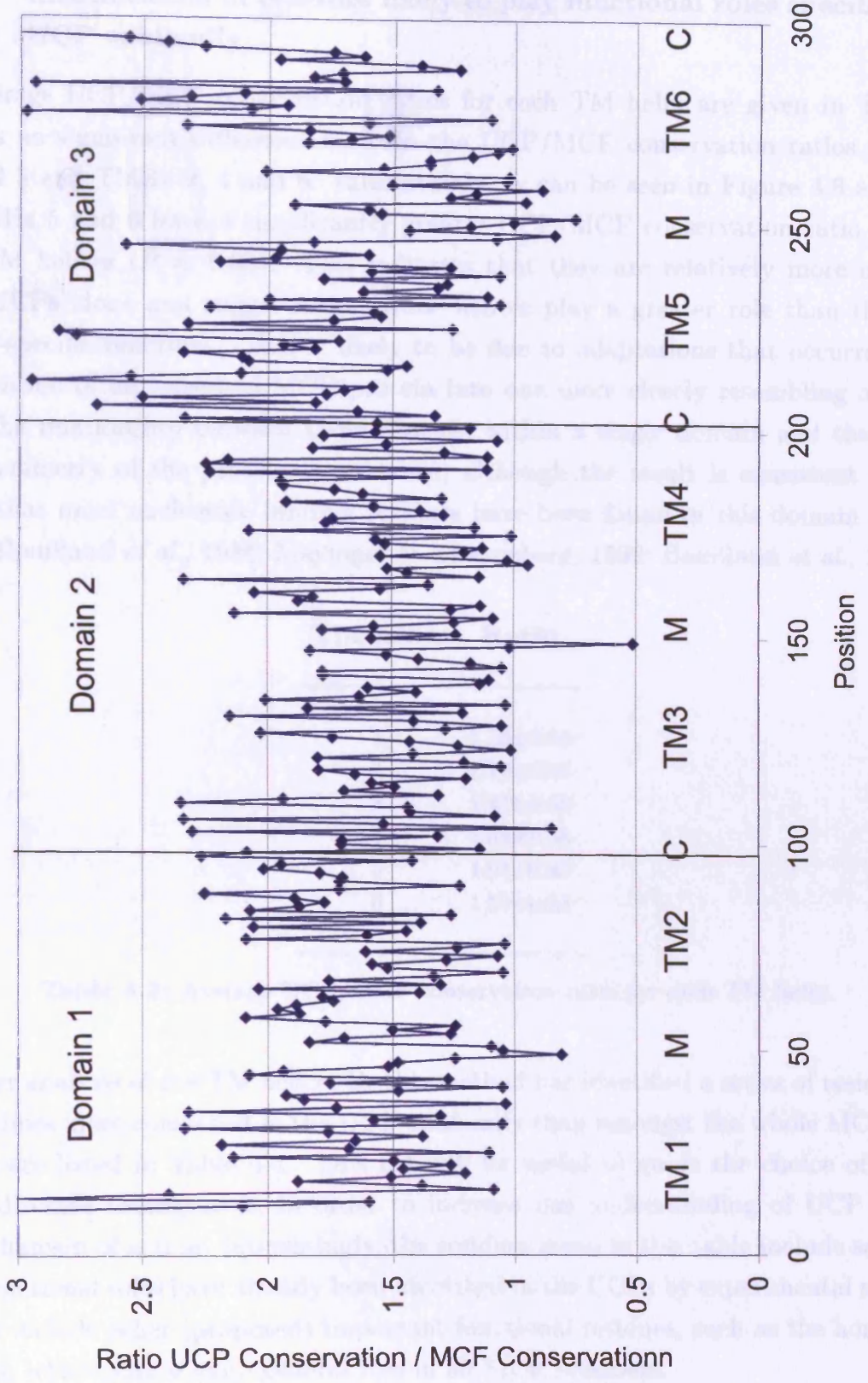


Figure 4.8: Ratio of UCP-derived conservation score to MCF-derived conservation score along the length of the UCP sequence. A ratio of 1 indicates equal conservation in the UCPs and the whole of the MCF, whereas a ratio of greater than 1 indicates greater conservation in the UCPs than in the rest of the family. Horizontal lines divide the UCP sequence into its 3 homologous domains. TM : transmembrane helix, C : cytosolic loop, M : matrical loop.

4.3.5.2 Identification of residues likely to play functional roles specific to the UCP subfamily

The average UCP/MCF conservation ratios for each TM helix are given in Table 4.3. There is no significant difference between the UCP/MCF conservation ratios of TMHs 1, 3 and 5 and TMHs 2, 4 and 6. Interestingly, as can be seen in Figure 4.8 and Table 4.3, TMHs 5 and 6 have a significantly greater UCP/MCF conservation ratio than the other TM helices ($P \approx 0.02$). This indicates that they are relatively more conserved in the UCPs alone and suggests that these helices play a greater role than the others in UCP-specific functions. This is likely to be due to adaptations that occurred in the specialisation of an ancestral MCF protein into one more closely resembling a modern UCP. The relationship between these changes within a single domain and the pseudo-3-fold symmetry of the protein is unknown, although the result is consistent with the finding that more nucleotide binding residues have been found in this domain than the others (Bouillaud *et al.*, 1986; Mayinger & Klingenberg, 1992; Bouillaud *et al.*, 1994).

TM helix	Ratio
1	1.56±0.38
2	1.72±0.35
3	1.47±0.35
4	1.63±0.38
5	1.91±0.48
6	1.92±0.55

Table 4.3: Average UCP/MCF conservation ratio for each TM helix.

Closer analysis of the TM helices by this method has identified a series of residues that are 2-3 times more conserved in the UCP-subfamily than amongst the whole MCF. These residues are listed in Table 4.4. This list will be useful to guide the choice of position for site-directed mutagenesis, in order to increase our understanding of UCP function and mechanism of action. Interestingly, the residues given in this table include several for which functional roles have already been identified in the UCPs by experimental means. It does not include other (proposed) important functional residues, such as the homologous arginines, which play a more general role in all MCF members.

W280 has been shown to be located in a water-filled cavity, and proposed to aid in binding of the nucleotide in the pore, via an aromatic interaction with its purine ring (Jezek *et al.*, 1998). Although it has not been investigated, the other aromatic residues in

Residue	Ratio	TM helix	UCP-specific role
L277	2.98	6	
M284	2.94	6	
C224	2.84	5	
F223	2.77	5	
W280	2.51	6	Interaction with purine ring
L219	2.34	5	
F31	2.33	1	
F274	2.33	6	
T226	2.32	5	
N281	2.31	6	
A88	2.25	2	
V192	2.25	4	
E190	2.23	4	pH control of nucleotide binding
S230	2.32	5	
D27	2.18	1	H^+ transport
Q82	2.17	2	
Q131	2.15	3	
C24	2.14	1	
H214	2.11	5	pH control of nucleotide binding
A218	2.11	1	

Table 4.4: 20 residues within the TM helices with the highest UCP/MCF conservation ratios, indicating functional roles specific to the UCPs. The final column identifies residues for which experimental evidence has already suggested functional roles (discussed in Chapter 2).

Table 4.4, F31, F223 and F274, may play a similar role. While the ADP/ATP translocase must also interact with nucleotides, since its role is in their transport it would be unlikely that similar residues would be involved. In particular, stabilising interactions provided by bulky groups like aromatic residues, which would likely hinder movement through a pore, seem unlikely. Hence a UCP-specific role would be expected for these residues.

In contrast, the inclusion of two cysteine residues, C24 and C224, in the list of residues that likely play a functional role in the UCPs is unexpected. Arechaga *et al.* (1993) have demonstrated that none of the 7 cysteine residues are essential for UCP function. Further experimental work is needed to explain this finding.

4.3.5.3 Identification of specific residues likely to have structural roles throughout the mitochondrial carrier family

Residues with general structural roles appear to occur with roughly equal frequency throughout the whole length of the sequence. They are shown in Figure 4.8 as being highly conserved in both the MCF and the UCPs. These residues, 10 of which are listed in Table 4.5, are likely to be found in buried positions, forming important helix-helix contacts.

Residue	MCF Conservation	TM helix	Angular error (°)
G175*	0.99	4	20
G123	0.99	3	60
G269*	0.99	6	20
G76*	0.96	2	20
G119	0.95	3	20
P32$^+$	0.95	1	120
P231$^+$	0.94	5	120
R276	0.91	6	40
P132$^+$	0.83	3	120
R83	0.80	2	40

Table 4.5: 10 TM residues with the highest conservation scores (across both the UCPs and MCF). These residues are likely to have general structural roles, such as TM helix packing. The angular error column shows the angle between the predicted buried vector and the position of the residue, giving an indication of whether it is found on the predicted buried face of the helix. * Indicates homologous glycines and $^+$ indicates homologous prolines.

Note that 5 out of the 10 residues in Table 4.5 are glycines, known to be both extremely common and important in TM helix packing (see Section 3.1.2, Chapter 3). The finding is consistent with the results of the prediction algorithm, since all but one of these glycines were found just 20° from the buried vector, suggesting they are found on the buried face of the helix. Strong evidence would be provided for a particular UCP model if it permitted these residues to be located in buried positions.

Despite their high conservation, the remaining residues in Table 4.5 are unlikely to be buried residues forming important helix-helix contacts. 3 of the residues are homologous prolines, residues known to have important structural roles due to their ability to kink helices. Hence, these residues are likely to have a structural role without necessarily being

buried, explaining their large angular errors given in Table 4.5. The final 2 residues in Table 4.5 are 2 of the homologous arginines. As discussed, there is some evidence to suggest that these residues play an important role in transport but it is not clear whether this role is direct. It is therefore not possible to determine whether their high conservation is due to a functional or a structural role throughout the family.

4.3.5.4 Importance of this work for UCP modelling

The implications of this work for UCP-model prediction are that the residues conserved over the entire MCF are likely to be more informative of structure than those conserved specifically amongst the UCPs. Hence, MCF-derived conservation scores will be used for the prediction of buried and lipid-tail-accessible UCP residues. Residues in the TM region that are specifically conserved in the UCP subfamily may indicate a pore-lining role, since this region would have required specialisation for H^+ transport and nucleotide/fatty acid regulation. These residues, identified above, should therefore be assigned a pore- lining position, if possible, during model optimisation.

Figure 4.9 shows residues with particularly high or low UCP/MCF conservation ratios on helical wheel representations for the UCP TM helices. TMHs 5 and 6 have a significantly greater UCP/MCF conservation ratio than the other TM helices, suggesting they play a greater role in UCP-specific functions. Mutagenesis may help to establish the role of these residues. However, there is no significant difference between the UCP/MCF conservation ratio of TMHs 1, 3 and 5 and TMHs 2, 4 and 6.

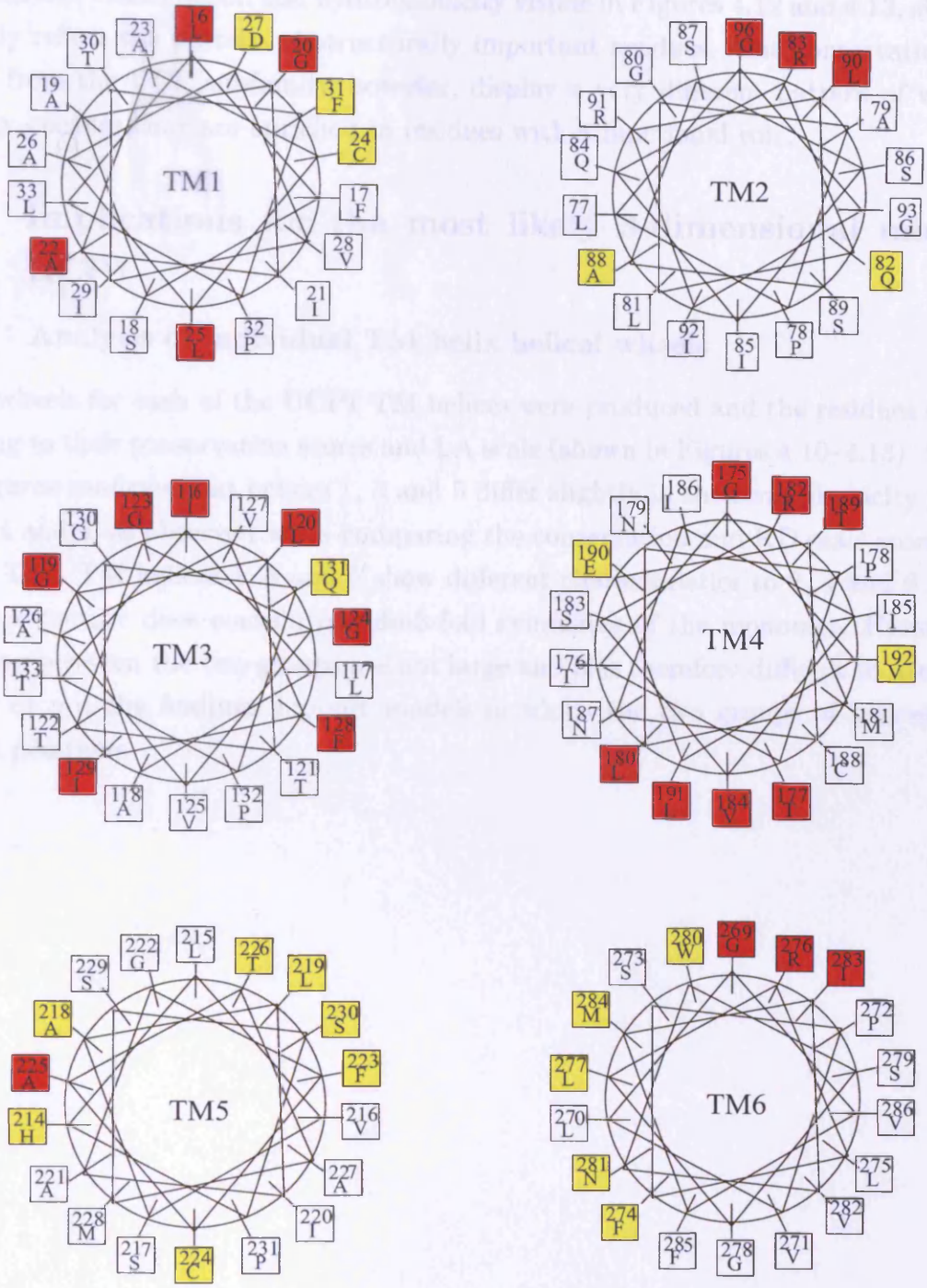


Figure 4.9: Helical wheel representations of the UCP TM helices, showing residues predicted to have a structural (red) or functional role (yellow) by analysis of their UCP/MCF conservation ratio.

The difference between UCP- and MCF-derived scores explains the correlation between family-derived conservation and hydrophobicity visible in Figures 4.12 and 4.13, since both primarily reflect the pattern of structurally important residues. The conservation scores derived from the UCP subfamily, however, display a very different pattern of variation, probably because they are enriched in residues with a functional role.

4.3.6 Implications for the most likely 3-dimensional model of UCP1

4.3.6.1 Analysis of individual TM helix helical wheels

Helical wheels for each of the UCP1 TM helices were produced and the residues coloured according to their conservation scores and LA scale (shown in Figures 4.10–4.13). Study of these figures confirms that helices 1, 3 and 5 differ slightly in their variphobicity patterns from 2, 4 and 6, as observed when comparing the conservation and KD scale scores of the helices. That TM helices 1, 3 and 5 show different characteristics to 2, 4 and 6 suggests that the structure does contain pseudo-3-fold symmetry of the monomer. However, the differences between the two groups are not large and it is therefore difficult to determined whether or not the findings support models in which the two groups of helices occupy different positions.

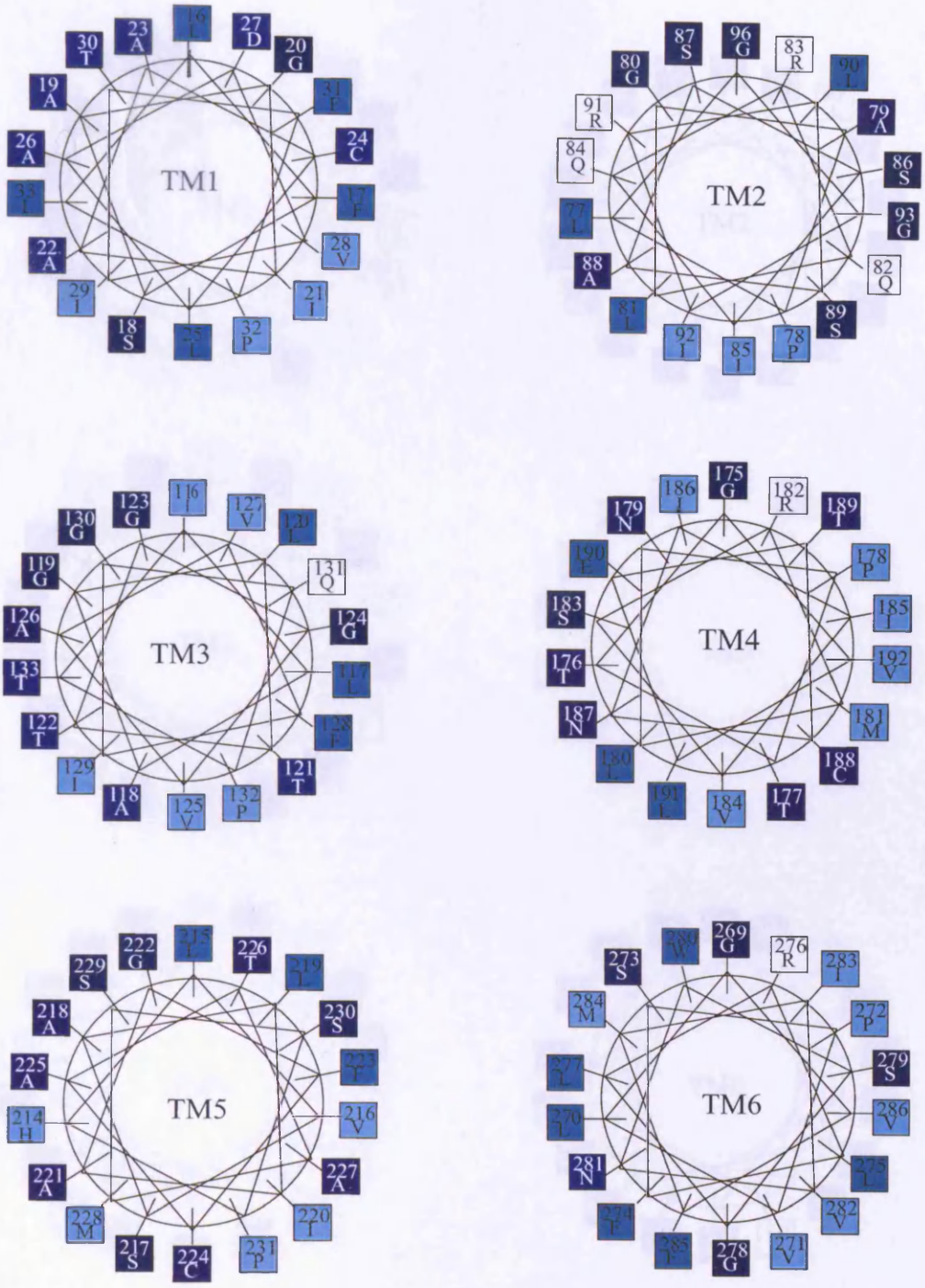


Figure 4.10: Helical wheels for the UCP TM helices, coloured by LA score. Colours run from white (most lipid-tail-accessible) to dark blue (most buried).

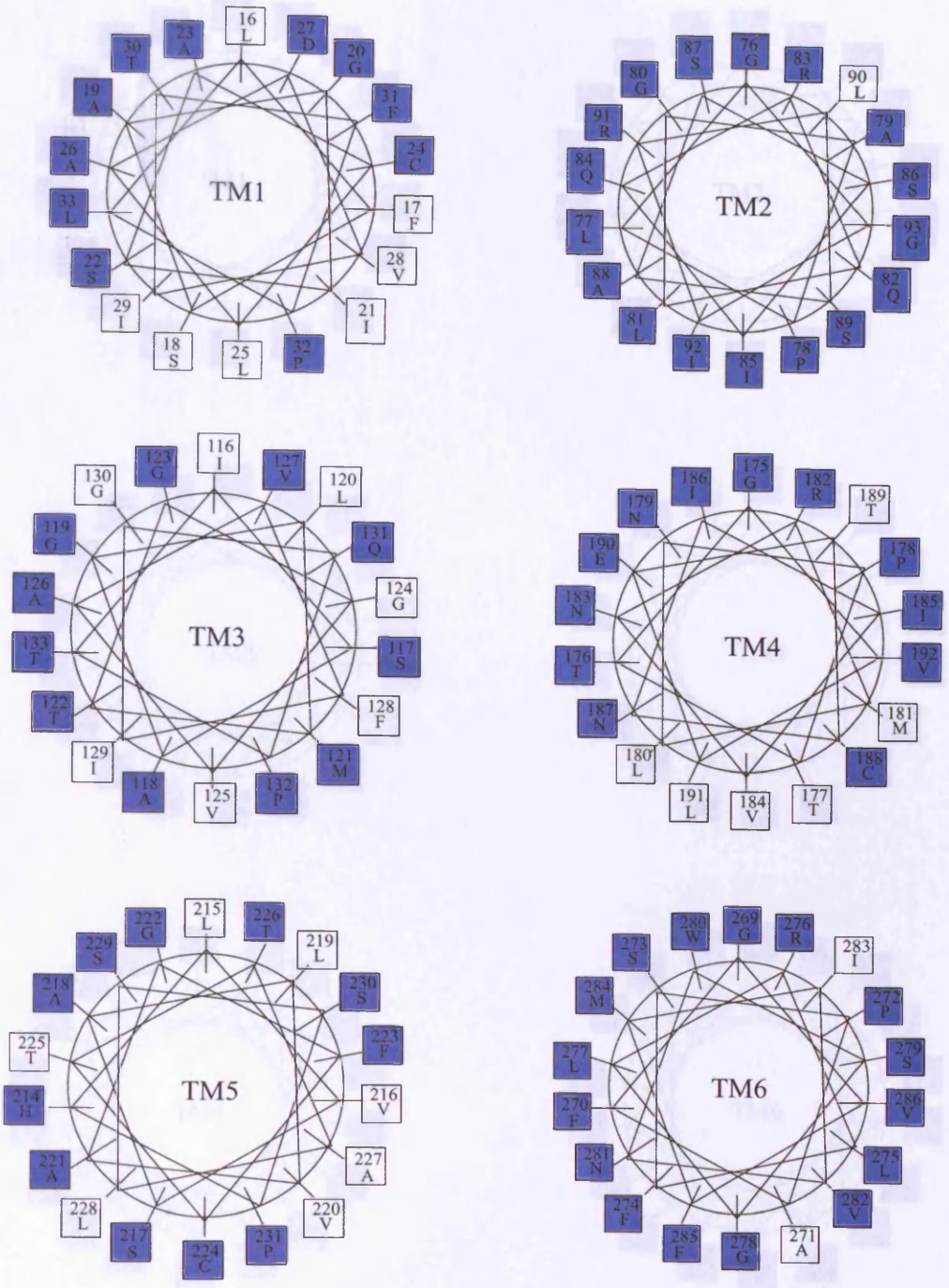


Figure 4.11: Helical wheels for the UCP TM helices, coloured by UCP-specific conservation score. Residues coloured blue show greater that 80% conservation over the UCP subfamily.

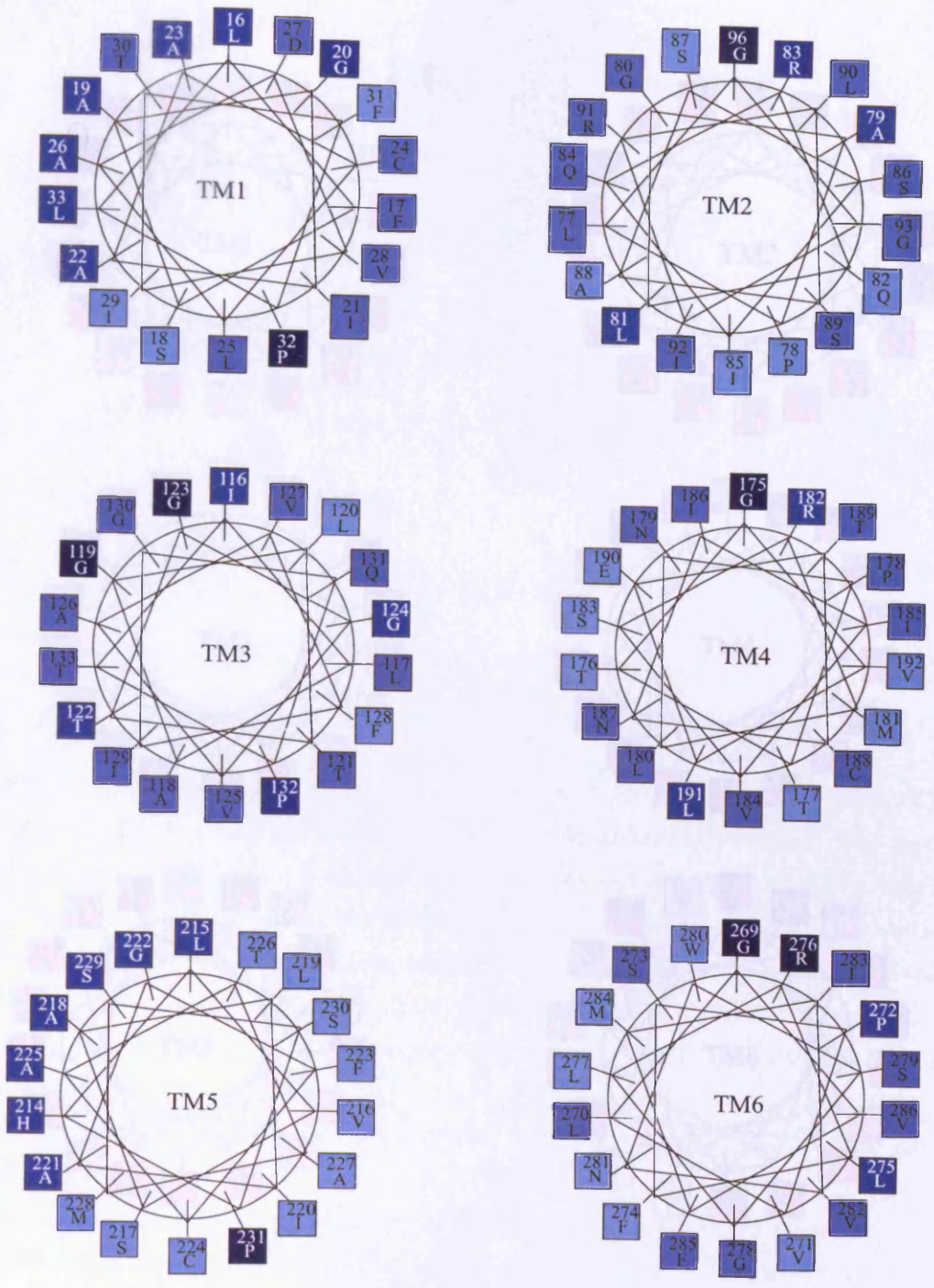


Figure 4.12: Helical wheels for the UCP TM helices, coloured by conservation score derived from the whole MCF. The key to the colouring of this figure is shown in Figure 4.14.

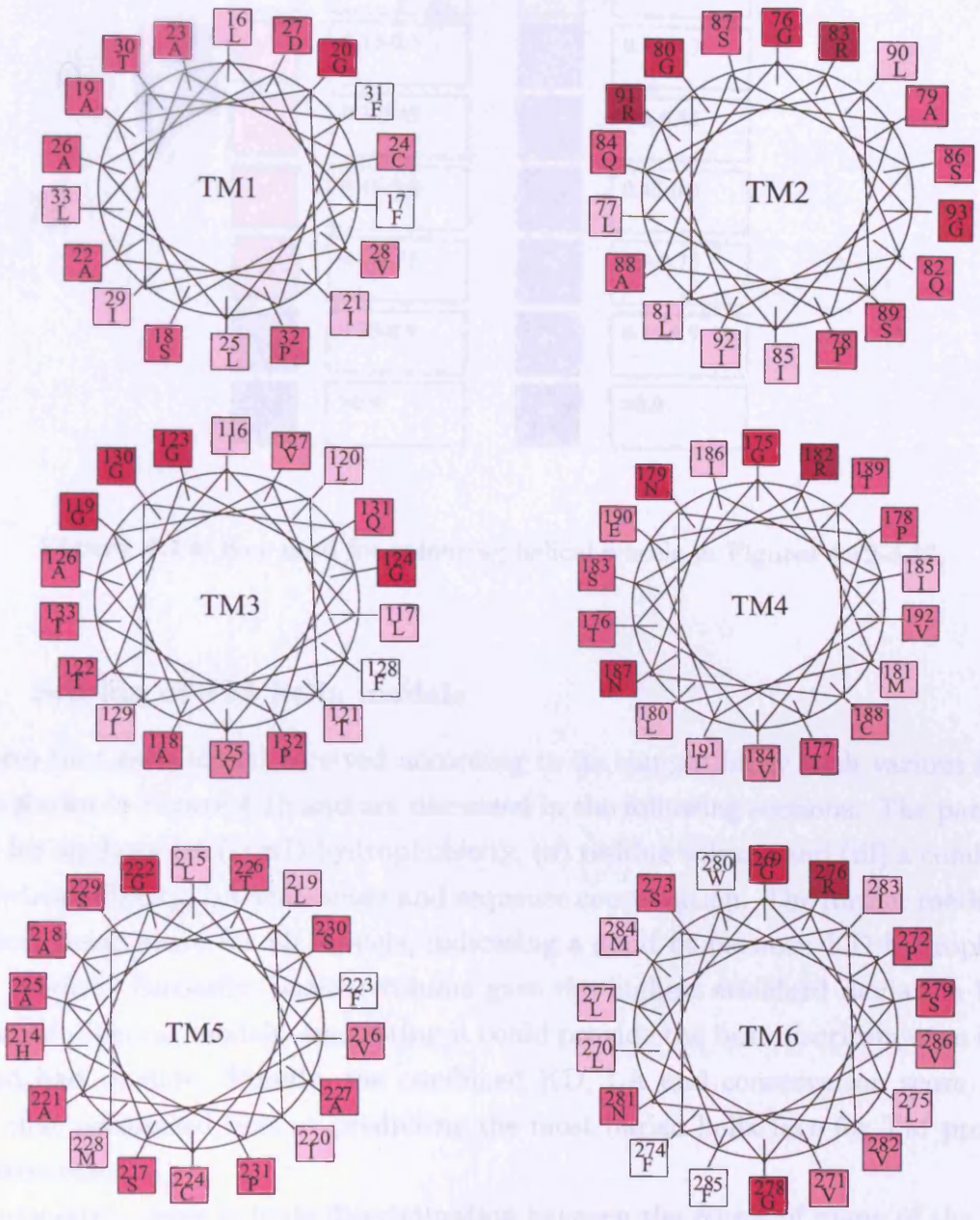


Figure 4.13: Helical wheels for the UCP TM helices, coloured by hydrophobicity on the White and Wimley scale (Wimley *et al.*, 1996; Jayasinghe *et al.*, 2001). The key to the colouring of this figure is shown in Figure 4.14.

Figure 4.14: Key used for colouring helical wheels in Figures 4.12-4.13.

4.3.6.2 Scoring of TM helix models

The scores that each model received according to its compatibility with various forms of data are shown in Figure 4.15 and are discussed in the following sections. The parameters selected for analysis are (i) KD hydrophobicity, (ii) residue volume and (iii) a combination of KD hydrophobicity, LA scale score and sequence conservation. The former method gave the highest mean score for all models, indicating a good fit between KD hydrophobicity and the models. Secondly, residue volume gave the highest standard deviation between the scores of different models, suggesting it could provide the best discrimination between good and bad models. Finally, the combined KD, LA and conservation score was the method that performed best at predicting the most buried helix face for TM proteins of known structure.

Unfortunately there is little discrimination between the scores of many of the models, particularly according to some parameters. This is likely to be due to the fact that some of the modelling assumptions, such as the straight, parallel nature of the helices, were perhaps too simplistic. If this were the case, even the correct model would not show a complete correspondence with the data since it could not account for these effects.

In general, the highest scoring models were those that are variations on Model 3, with the dimeric unit consisting of 2 rings of 6 TM helices, each surrounding a pore. In addition, a model in which the monomer-monomer interface is formed by helices 2, 3 and

4 seems to be most compatible with the data. These findings are discussed in more detail in the following sections.

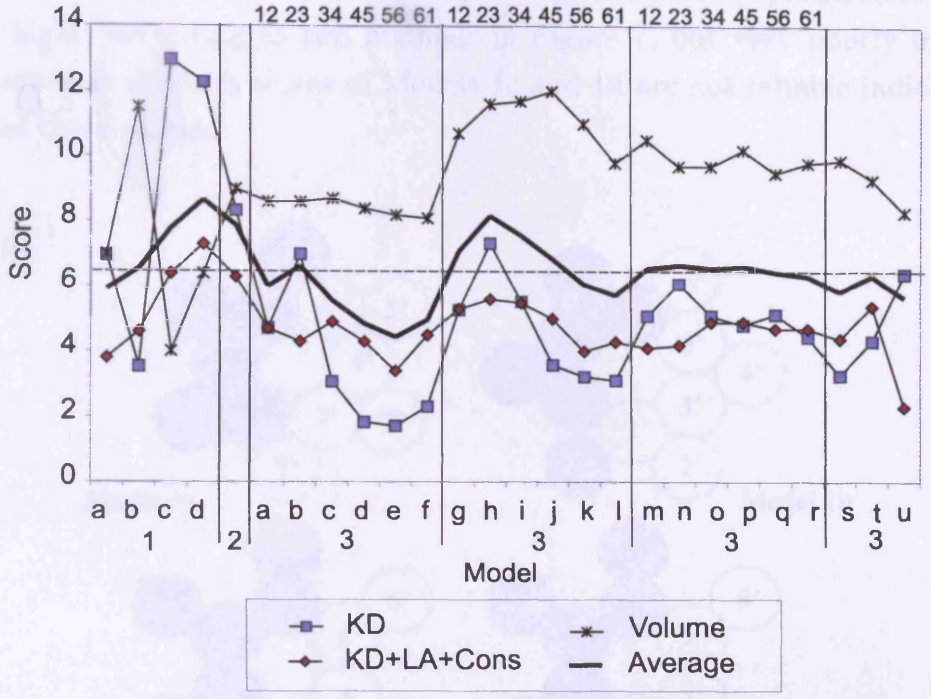


Figure 4.15: Average scores for each model, according to their compatibility with several forms of sequence data. 'Average' indicates the average of the other methods shown. In each of the variants of Model3a-r the monomer-monomer interface is formed by the helices shown above. The structures of Models 1a-1d are shown in Figure 4.16. Model 2 consists of a ring of 12 TM helices surrounding a pore. Models 3a-3u are illustrated in Figure 4.17. Models 3a-r are dimeric and 3s-u are monomeric. In Models 3a-3f and 3s the odd and even numbered helices are found in equivalent positions. In Models 3g-3l and 3t the odd numbered helices are found in more buried positions than the even. In Models 3m-3r and 3u the even numbered helices are found in more buried positions than the odd.

Model 1 Figure 4.16 shows schematically the variations of Model 1, referred to as Models 1a, 1b, 1c and 1d. While Models 1a and 1b obtain very low scores, they show a compact arrangement of TM helices, similar to that seen in other TM proteins (Chapter 3, Section 3.3.4). In contrast, the most loose-packed Models, 1c and 1d, are less similar to proteins of known structure. As can be seen in Figures 3.6–3.8 in Chapter 3, the majority of TM proteins tend to have more compact arrangements of TM helices, similar to that seen in Models 1a and 1b. However, the models showing a 'loose packing' of the helices (Models 1c and 1d), cannot be discounted for this reason, because the arrangement of helices is

similar to that seen in ATP synthase subunit C (Girvin *et al.*, 1998). Interestingly, this protein is also an H^+ channel.

The scores of Models 1a-d are very sensitive to the scoring parameters used: they score very highly according to two methods in Figure 7, but very poorly in the other. This suggests that the high scores of Models 1c and 1d are not reliable indicators of the likelihood of these models.

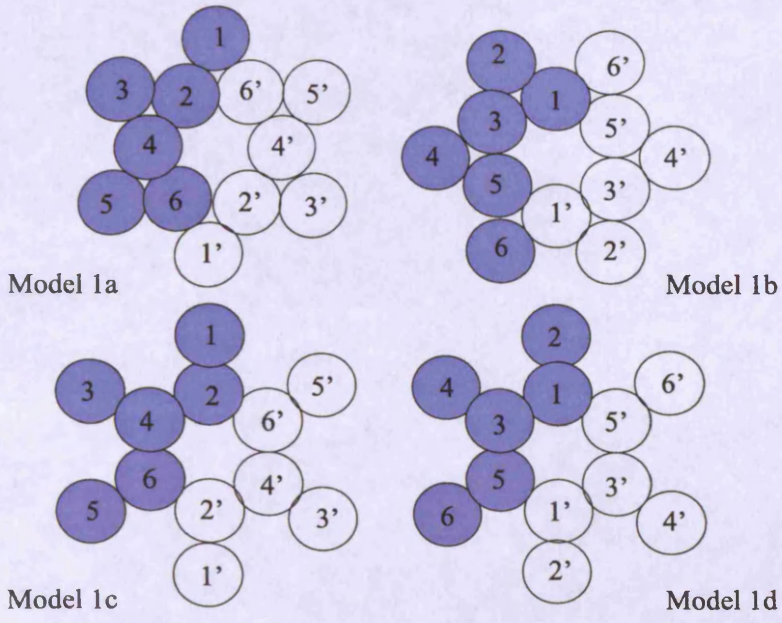


Figure 4.16: Models 1a, 1b, 1c and 1d: Alternative arrangements of uncoupling protein transmembrane helices for Model 1. Each helix, represented as a circle, has a diameter of 10Å.

Models 1a and 1c would permit a pore-lining location for the homologous arginines found on TM helices 2, 4 and 6, consistent with their proposed role in binding of the nucleotide phosphate (Modriansky *et al.*, 1997). However, the presence of these residues in virtually all members of the MCF, most of which do not bind nucleotides, argues against such a direct role (Echtay *et al.*, 2001a). Hence this data cannot be used to select between Models 1a and 1c, in which TMHs 2, 4 and 6 are pore-lining and Models 1b and 1d, in which they are located peripherally.

In summary, while Models 1c and 1d score the most highly, Models 1a and 1b show a compact arrangements of TM helices, similar to that seen in other TM proteins. However, the scores are very dependent upon the scoring parameters used.

Model 2 Model 2 has obtained a relatively high score compared to other arrangements, suggesting a good correlation with the available data. However, this model has a very different structure to the other TM proteins in the dataset of Chapter 3. It also seems unlikely that such a large pore would have evolved, even in order to transport the largest substrates of the family such as ATP. It would be particularly difficult for the UCPs to maintain H^+ specificity through such a large pore. These factors suggest that overall, Model 2 is an unlikely candidate for UCP structure.

Model 3 The variations of Model 3 are shown in Figure 4.17. As shown in this figure, they can be grouped into 3 classes. These are those in which the even and odd-numbered helices occupy equivalent positions, (Models 3a-f) those in which TMHs 2, 4 and 6 are more peripheral (Models 3g-l) and those in which TMHs 1, 3 and 5 are more peripheral (Models 3m-r). Monomeric forms of Model 3 are also considered (Model 3s-u).

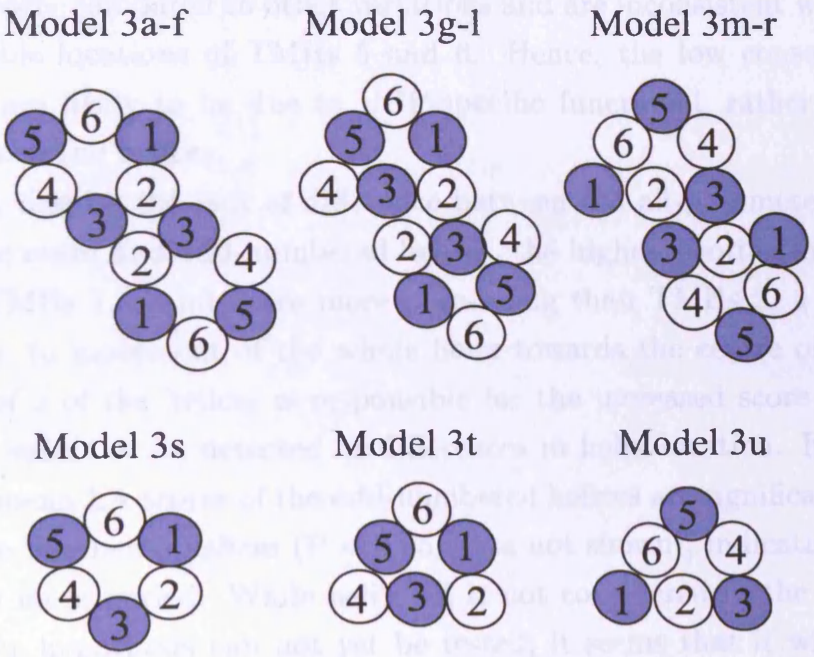


Figure 4.17: Models 3a-u: Alternative arrangements of uncoupling protein transmembrane helices. Models 3b, 3h and 3n are shown as representative examples of each class of dimeric models, although any 2 sequential helices may form the monomer-monomer interface for each class. Each helix, represented as a circle, has a diameter of 10Å. TMHs 1, 3 and 5 are shaded to highlight their different positions in each group of models. Model 3s-u are a monomeric forms of Models 3a-f.

Models 3s, 3t and 3u are contrary to the evidence that suggests that the UCPs are

functional as a dimer. However, they do obtain scores relatively similar to most other Model 3 variants, suggesting that the oligomeric structure of the UCPs may merit further investigation. All forms of Model 3 assume that the weak experimental evidence for a single-pore arrangement of the UCP dimer is incorrect (Schroers *et al.*, 1998; Huang *et al.*, 2001). They show a pore size similar to those observed for TM proteins of known structure in Chapter 3. Interestingly, the structure of these models would allow independent folding of each monomer, before dimerisation. On the other hand, it is harder to propose a plausible folding mechanism for the other models, in which the nascent chains must first dimerise before undergoing simultaneous folding. Crucially, Model 3 also includes all of the highest scoring models, making it appear a likely candidate for UCP structure.

Model 3 is the only one which could account for the greater UCP/MCF conservation ratio observed for TMHs 5 and 6, by permitting them to form the monomer-monomer interface. However, this theory requires that the UCPs are the only members of the family to dimerise in this way. In addition, the models in which this is the case (Models 3e, 3k and 3q) score poorly compared to other variations and are inconsistent with the predicted lipid-tail-accessible locations of TMHs 5 and 6. Hence, the low conservation ratios of TMHs 5 and 6 are likely to be due to UCP-specific functional, rather than structural, differences between the helices.

Interestingly, despite the lack of difference between the all-parameter combined location scores of the even- and odd-numbered helices, the highest scoring models are Models 3g-l, in which TMHs 1, 3 and 5 are more pore-lining than TMHs 2, 4 and 6. Perhaps, as an alternative to movement of the whole helix towards the centre of the bundle, the relative tilting of 3 of the helices is responsible for the increased score of these models, while being too subtle to be detected as differences in helix location. In support of this hypothesis, the mean LA scores of the odd-numbered helices are significantly greater than those of the even-numbered helices ($P < 0.05$, data not shown), indicating that helices 1, 3 and 5 may be more buried. While helix tilt is not considered in the current strategy, and therefore the hypothesis can not yet be tested, it seems that it will be valuable to include this factor in future methods.

The second highest scoring of all models is Model 3h. This model is consistently the highest scoring of the variations of Model 3. In this model, the monomer-monomer interface is formed by TMHs 2 and 3. In addition, the high scores of Models 3g and 3i suggest that, if Model 3h is correct, while TMHs 2 and 3 play the major role in the interface, a smaller contribution in made by TMHs 1 and 4. This finding is very consistent with the predicted buried location of TMHs 2 and 3, relative to the others. It is also interesting that the use of KD hydrophobicity as a scoring parameter detected increased scores for the other models with interfaces formed by TMHs 2 and 3, Models 3b

and 3n (Figure 4.15). The likely importance of TMHs 2 and 3 in the monomer-monomer interface has also been suggested by Nelson & Douglas (1993).

In conclusion, several very highly scoring models are variations on Model 3, with monomeric or dimeric units consisting of rings of 6 TM helices surrounding a pore. However, the highest scoring variations of Model 3 are inconsistent with the prediction that the even- and odd-numbered helices are predicted to occupy equivalent positions, due to their similar all-parameter combined scores. This suggests that some combination of the two arrangements may be the correct one: in which the helices are tilted to allow the even numbered helices to be slightly more buried that the others, perhaps only at one end of the pore. That TMHs 2 and 3 form the monomer-monomer interface seems likely. Overall it is difficult to draw strong conclusions about the correct model from these data since the scores are relatively similar and two contradictory models (Models 1d and 3h) receive similarly high scores.

4.3.6.3 Analysis of the experimental evidence

There are few, if any, pieces of experimental evidence which unequivocally restrict the location of a residue to a particular environment. Firstly, some of the data can easily be discounted. For example, while there have been suggestions of an ionic bond between R91 and E190 (Echtay *et al.*, 2001a), helical wheel representations show that this bond is impossible if equivalent positions are to be observed for homologous helices, in accordance with pseudo-3-fold symmetry. This constraint can therefore not be used in modelling.

The second problem derives from the inability to determine whether the loss of function (eg H^+ transport or nucleotide regulation) that occurs on mutation is caused directly by the loss of a participating residue, or indirectly due to disruption of the native conformation. For example, there have been suggestions that D27 (Klingenberg & Echtay, 2001) and the homologous arginines (Echtay *et al.*, 2001a) have pore lining positions, but there is also strong evidence against a direct role in transport (Echtay *et al.*, 2000a; Urbankova *et al.*, 2003; Echtay *et al.*, 2001a; Modriansky *et al.*, 1997), failing to confirm this location. Similarly, it is not known whether the role of H214 in regulation of nucleotide binding requires the direct positioning of H214 itself within the pore. Consequently, little weight can be attached to the degree of correspondence between particular models and experimentally derived information.

With this in mind, the degree of correspondence between each of the optimised models and the experimental data has been assessed, and is summarised in Table 4.6. Models 2 and 3 are the most consistent with the experimental data, while Models 1b and 1d are the least consistent. However, the experimental data are not conclusive enough to be

Mutant	1a	1b	1c	1d	2	3	UCP
D27 lines transport pathway	-	+	-	+	+	+	+
E190- pH sensitivity	+	-	+	-	+	+	+
H214- pH sensitivity	-	+	-	+	+	+	+
R83/R182/R267 line H^+ transport path	+	-	+	-	+	+	+
R91/E190 H bond	-	-	-	-	-	-	-
W280 in water-filled cavity	+	-	+	-	+	+	-*
C24/188/224 lipid-tail-accessible	+	-	+	+	+	+	+
Number correct/7	4	2	4	3	6	6	5

Table 4.6: Comparison of the degree of agreement between each of the optimised models and the experimental data. The experimental data is described in Chapter 2 and summarised in Table 4.1. The column entitled UCP indicates the agreement of the UCP homology model, described in Section 4.4.1, with the data. A + indicates that a model is consistent with a particular piece of experimental evidence and a - indicates that they are inconsistent. Notes: 1. The pore-lining location of all three homologous arginines was considered to be consistent with one piece of experimental evidence. 2. The R91/E190 H bond is impossible for all models that show pseudo-3-fold symmetry. 3. * indicates that the lack of agreement between this data and the homology model is likely to be due to problems with the sequence alignment used.

confident of a single correct model.

4.3.6.4 Conclusions

It is not possible to select between the models with high confidence using any single method in isolation: experimental data, predictions of helix location or scoring according to compatibility with sequence data. However, after study of the combined results of these analyses, summarised in Tables 4.7–4.9, it can be concluded that a model similar to Model 3h but with tilted helices is perhaps the most likely, since it:

- obtains the second highest score for compatibility with the sequence data
- is consistent with 6 out of 7 pieces of experimental data
- is able to account for the likely buried position of TMHs 2 and 3, within the monomer-monomer interface, as predicted by their all-parameter combined score

- is able to reconcile the apparently contradictory predictions of helix location obtained using the all-parameter combined score and LA score, because helix tilting may permit different relative positions of helices at different heights in the membrane

This model consists of two bundles of 6 TM helices, each around a pore. The helices may be tilted and the monomer-monomer interface is likely to be formed mainly by TMHs 2 and 3, with smaller contributions from TMHs 1 and 4. While Model 1d scored slightly higher than the selected Model 3h according to compatibility with some forms of sequence-based data, other sources of evidence suggest that this model is considerably less likely than Model 3h.

Evidence for Models 1a, 1c Helices 2, 4 and 6 lining the pore	**Evidence for Models 1b, 1d** Helices 1, 3 and 5 lining the pore
Models 1a and 1c score most highly for compatibility with some forms of data	Models 1b and 1d score most highly for compatibility with other forms of data
Model 1a is supported by 4 pieces of experimental evidence (Experimental data favours models with TMHs 2, 4 and 6 lining the pore, although this evidence is unreliable)	Model 1b is supported by only 2 pieces of experimental evidence
The kinking caused by the conserved prolines found in TMHs 1, 3 and 5 would be less disruptive to helix packing if they were peripherally, as is seen in ATP synthase subunit C	TMHs 1, 3 and 5 contain a conserved proline that may facilitate a conformational change associated with transport

Table 4.7: Evidence for and against the pore-lining location of UCP TM helices 1, 3 and 5 (Models 1b and 1d) and 2, 4 and 6 (Models 1a and 1c). These models are illustrated in Figure 4.16.

Evidence for Model 2	Evidence against Model 2
Consistent with all 7 pieces of experimental data	Significantly greater conservation ratio of TMHs 5 and 6
	Model 2 shoes a dissimilar proportion of pore-lining residues to TM proteins of known structure

Table 4.8: Evidence for and against Model 2. Model 2 consists of a ring of 12 TM helices around a pore.

Evidence for Model 3	Evidence against Model 3
Models 3h/i score very highly for compatibility with various forms of data	Inconsistent with weak evidence suggesting a single pore per dimer
Consistent with 6 pieces of experimental evidence	Significantly greater conservation ratio of TMHs 5 and 6 consistent only with Models 3e, 3k and 3q
Similar pore size to TM proteins of known structure	
Assembly of monomeric units easy for these models	

Table 4.9: Evidence for and against Model 3. For descriptions of the variations of Model 3, see Section 4.3.6.2.

4.4 Discussion

4.4.1 Comparison of the UCP model with the actual structure of the adenine nucleotide carrier

While this work was being completed, the structure of the adenine nucleotide carrier (ANT), another member of the mitochondrial carrier protein family, was solved by X-ray crystallography at a resolution of 2.2Å (PDB code 10KC, R factor 0.22) (Pebay-Peyroula *et al.*, 2003). The structure is shown in Figures 4.18 and 4.19. This protein would be expected to show the same fold as the UCPs, since they are evolutionarily related (the ANT shows 20% sequence identity and 25% similarity to UCP1). Hence its structure can be used to assess the accuracy of the model proposed in this chapter for the UCPs.

The protein does indeed consist of a pseudo-3-fold symmetric ring of 6 TM helices around a pore with a diameter of 15-20Å, consistent with the predictions made in this chapter. However, the protein exists as a monomer in the crystal, so its putative dimeric status in vivo needs further investigation. It is, therefore, not yet possible to establish whether or not the prediction that the monomer-monomer interface is formed by TMHs 2 and 3 is correct.

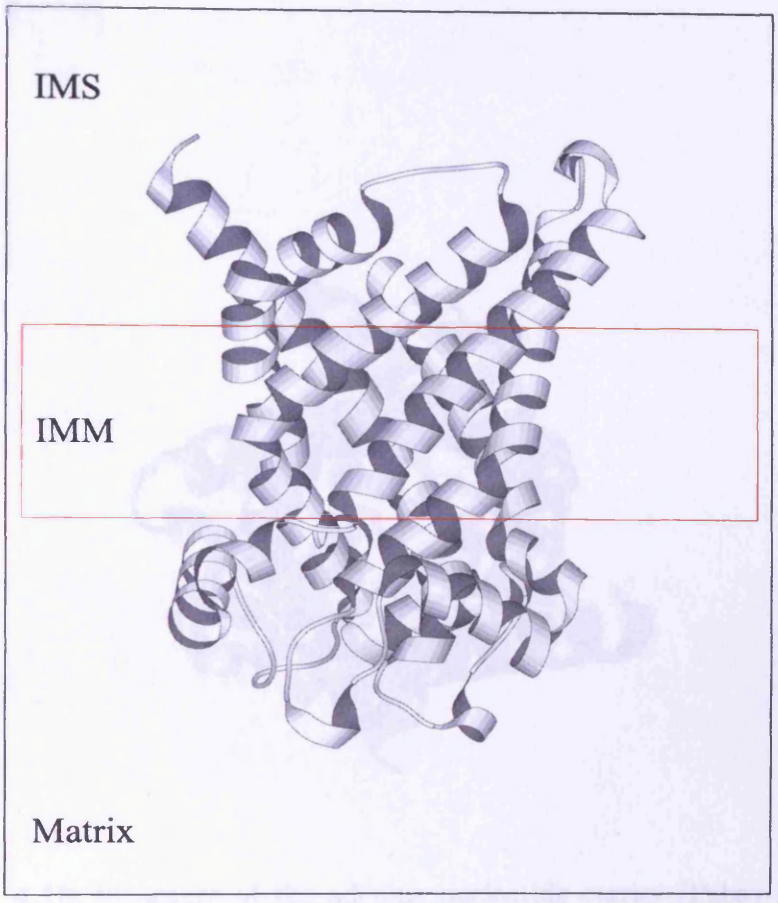


Figure 4.18: Structure of the adenine nucleotide carrier (Pebay-Peyroula *et al.*, 2003) in a view perpendicular to the membrane normal. IMS: Intermembrane space; IMM: Inner mitochondrial membrane; Matrix: Mitochondrial matrix. The red box shows the location of the membrane lipid-tail-spanning and head-group-spanning regions, as defined by PSlice (see Chapter 3). This figure was produced using MolScript.

Figure 4.19: Structure of the adenine nucleotide carrier (Pebay-Peyroula *et al.*, 2003) in a view along the membrane normal. This figure was produced using MolScript.

The performance of the prediction of buried helix faces is summarised in Figure 4.20. The best performance in proteins of known structure is based upon a combination of KD, LA and conservation scores with one residue excluded from each end of all helices. However, this method performs relatively poorly for the UCPs, with an average angular error of 71°. The best prediction for the UCP helices is made by KD and LA score, giving an angular error of 25°. That both of these angular errors are considerably greater than the average for proteins of known structure (13°) suggests that the prediction was more difficult for the UCPs than for proteins on average (assuming that the dataset used was a representative set of all proteins).

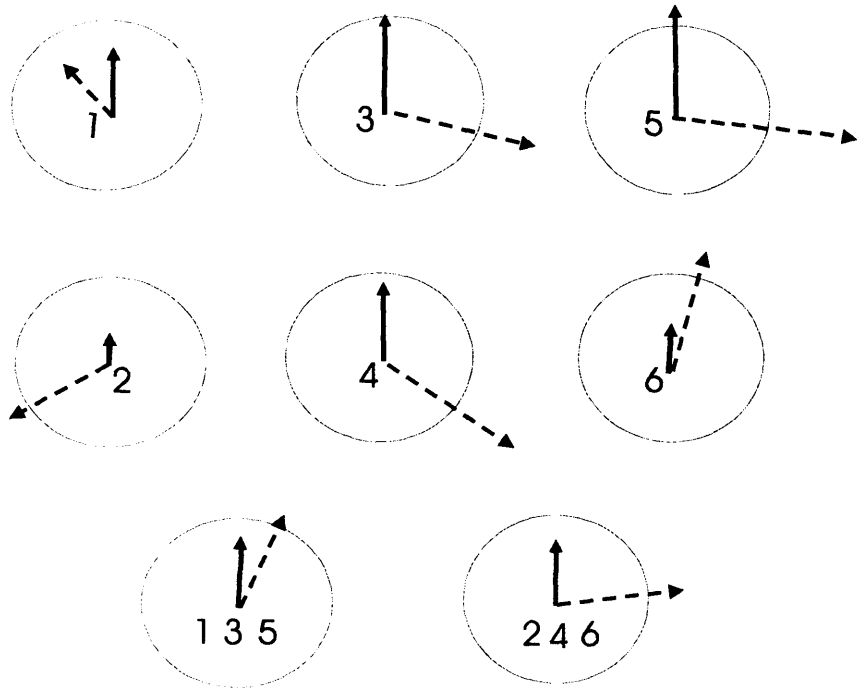


Figure 4.20: Schematic diagram showing the relative positions of the predicted (solid arrows) and actual (dashed arrows) buried faces of the UCP TM helices. The prediction was made, in this case, using KD hydrophobicity, LA score and conservation for the central helix residues. When the helices are considered independently (top 6 helical wheels), the average angular error between the predicted and actual vectors is 62°. When 3-fold symmetry of homologous helices is preserved (lower 2 helical wheels) the average angular error is 54°.

The accuracy of the prediction for the UCPs was reduced when pseudo-3-fold symmetry was not enforced (Figure 4.20). (This was achieved by predicting each helix individually, rather than combining the features of helices 1, 3 and 5 and of helices 2, 4 and

6, to give a single prediction for each group). Using this method, the average angular error was increased from 25 to 69°, for a prediction based on KD and LA score in the helix centre. This result illustrates the value of including in a predictive method specific biological knowledge about the protein of interest, such as that concerning symmetry.

The problems that have been encountered with the UCP modelling are likely to be due to the high tilt angles of all of the helices, and the kinks caused by the conserved prolines in TMHs 1, 3 and 5. For simplicity, the current method models TM helices as both straight and parallel to the membrane normal. As demonstrated by the high accuracy of the method on the proteins of known structure (mean angular error 13°), the assumptions made were valid based on the available data. Unfortunately, however, the adenine nucleotide carrier shows more helix kinking and tilting than most of the previously available structures, and is therefore modelled less accurately.

The assumption of straight, parallel helices allowed a single face of each helix to be identified as buried, as can be seen in Figure 4.21(A) as a vertical stripe of residues. When this is not the case in the actual structure, even the closest model will not obtain a very high score for compatibility with the sequence data. This is observed in Figure 4.21(B), where several scattered groups of residues are buried, rather than the predicted vertical stripe. It is also shown in Figure 4.20, where, despite pseudo-3-fold symmetry of structure, accessibility vector sums do not generally identify homologous buried faces of the helices. It is therefore not surprising that there was little differentiation between correct and incorrect models throughout this work. However, it was assumed that for proteins where kinking or tilting occurs, the correct model will still be identified because it will show a better fit than the other models with the data.

The tilt of the UCP helices creates a pore that is funnel-shaped, with the widest part at the cytosolic end. At the cytosolic side of the membrane the even and odd-numbered helices seem to contribute roughly evenly to the pore, very similar to Model 3s. At the matrix side, however, the pore is almost entirely closed due to the protrusion of the ends of TMHs 1, 3 and 5, more similar to model 3u. The protrusion of helices 1, 3 and 5 into the pore is made possible because homologous prolines (P32, P132 and P231) cause these helices to kink by 50-60° inwards. This narrowing of the pore is likely to form part of the gate that gives the protein its observed transporter-like, rather than channel-like, properties (Arechaga *et al.*, 2001), as described in Chapter 2.

Now that the structure of the ANT is known, it can be used as a template for homology modelling, to obtain a high resolution structural model of the UCPs. The sequences of the ANT, UCP1, UCP2, UCP3, and two other members of the mitochondrial carrier protein family, the deoxynucleotide carrier and the citrate transporter, were aligned using ClustalW (Thompson *et al.*, 1994) and a homology model was produced using the SWISS-

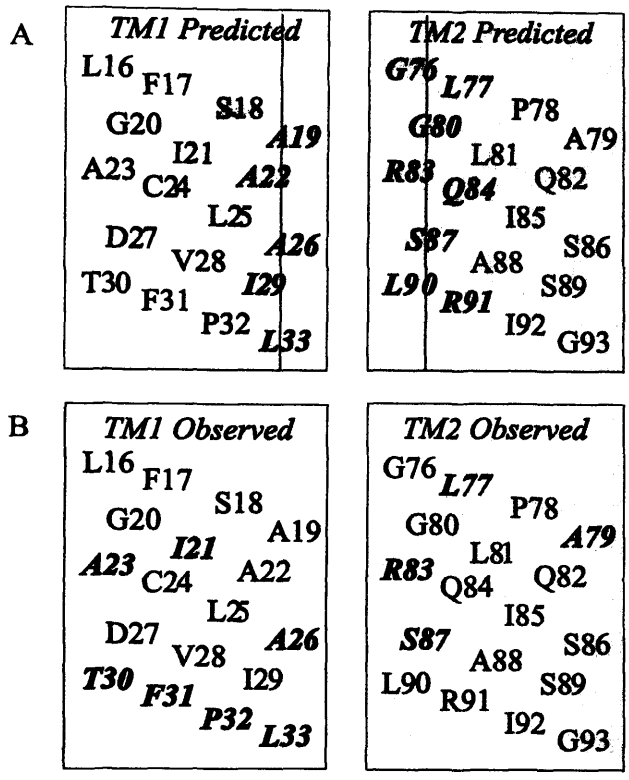


Figure 4.21: Helical nets of UCP TM helices 1 and 2, showing (A) the predicted buried face and (B) the actual buried residues, as calculated from the structure of the adenine nucleotide carrier (Pebay-Peyroula *et al.*, 2003). Buried residues are shown in bold and italics. Actual buried residues were taken as those with less than 10% relative accessible surface area.

MODEL server (Schwede *et al.*, 2003). The alignment is given in Figure 4.22.

Despite a sequence identity of only 20% between UCP1 and the ANT, ClustalW was able to align the sequences giving regions of similarity spread roughly evenly throughout their whole length. Of the 307 residues in UCP1, 20 were totally conserved across all 6 aligned sequences (shown in Figure 4.22). Almost identical levels of similarity are found within the TM helices and the extra-membrane loops, suggesting that both regions have important roles. (31% of TM residues were identical to all other residues at that alignment position, or showed only conservative or semi-conservative substitutions, compared to 32% of non-membrane-spanning residues). The majority of the insertions and deletions are found in the loops connecting the TM helices. The only exceptions to this are a two-residue deletion in TMH5 and a single-residue deletion in TMH6 of the ANT relative to the UCPs. Strongly conserved regions at either end of these helices prevent the alignment being adjusted to place these deletions in loop regions, and hence the model is only an

approximate one. However, it remains valuable, since the ability to assess the degree of correspondence between the UCP model and the experimental data is unlikely to be greatly affected.

As shown in Table 4.6, the homology model is consistent with virtually all of the experimental evidence concerning the likely role and location of the UCP residues. As suggested by the experimental data, the functional residues D27, E190, R83 and R182 are all found in a pore-lining position in the UCP model. Conversely, the cysteine residues at positions 24, 188 and 224 are all found in lipid-tail-accessible positions, as suggested by the lack of effect on UCP function caused by their mutagenesis (Arechaga *et al.*, 1993). While quenching studies suggested that W280 is found in a water-filled cavity (Jezek *et al.*, 1998), the model indicates that this residue is likely to be lipid-tail-accessible. Similarly, R276 is lipid-tail-accessible in the UCP model, despite an experimentally determined role in UCP function indicating either a buried or pore-lining position (Echtay *et al.*, 2001a; Modriansky *et al.*, 1997). However, these problems may be explained by the fact that the positions of W280 and R276 in the model are likely to have been affected by the deletion at position 277. In support of this, the other homologous arginines are both pore-lining. There is no evidence from the UCP homology model to suggest that the proposed R91/E190 ionic bond is formed. Therefore, in agreement with the results of Chapter 3, the presence of opposing charges in adjacent helices does not necessarily indicate that these residues will interact or even be located close to one another in the 3-dimensional structure.

It can be concluded that, while the experimental data is generally consistent with the actual UCP structure, it is very difficult to use this information to constrain the number of possible models. This is because the data is likely to be consistent with a number of alternative structures. If mutagenesis data is to be used to constrain the number of models, a large number of mutants at carefully selected positions will be needed, and careful interpretation of the result will be required. Often, any value of mutagenesis and related experimental techniques in modelling is likely to be restricted to verifying the likelihood of a model proposed by another method.

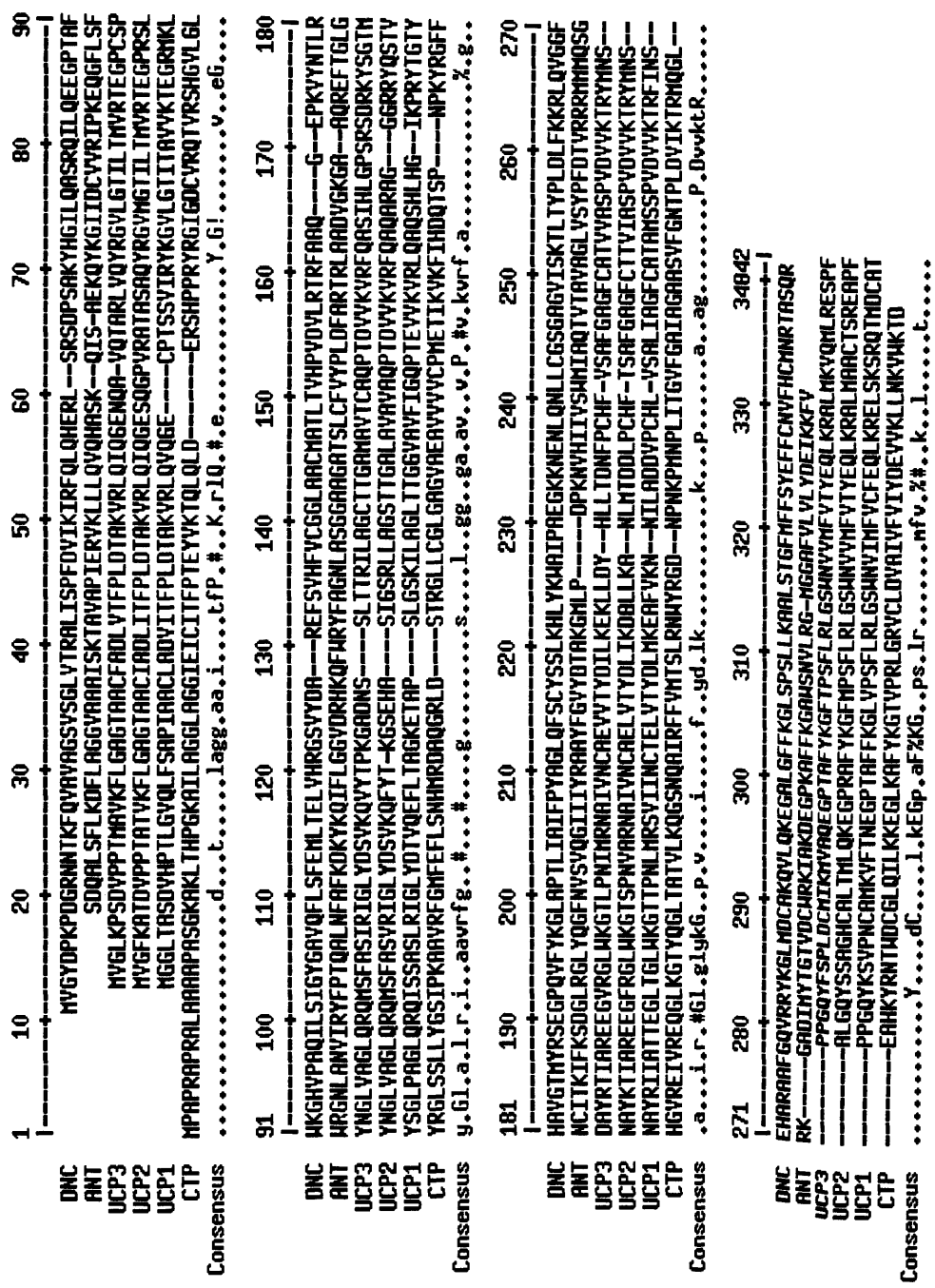


Figure 4.22: Multiple sequence alignment of the adeninine nucleotide carrier (ANT) and several other members of the mitochondrial carrier protein family, used to generate the homology model for UCP1. CTP: Citrate transport protein; DNC: Deoxynucleotide carrier. The coloured alignment was generated using MultiAln (Corpet F, 1988). Red and blue indicate respectively positions with greater that 90% and 50% identity.

4.4.2 Conclusions

The value of the approach taken to TM protein modelling in this work lies in the potential to produce models in the absence of any structural information for the family concerned, and without relying upon the use of soluble protein structural information. The method has been shown to be effective at detecting buried helix faces in TM proteins of known structure, and will be of use for this purpose for all membrane proteins where structural information is desired. The method has also enabled the UCP models to be ranked in order of likelihood, permitting selection a of single model which fits the available data most closely. Models proposed by this method will be a useful starting point in experimental studies aimed at increasing our understanding of the structure and function of any protein family, particularly by targeting cross-linking and mutagenesis studies. However, the model proposed for the UCPs is highly tentative. The model selected did not score significantly higher than other models by any one method used here. Hence a single technique has not yet been identified that is able to select one model above the others in an automatable way, at least for the UCPs.

Since the structure of a related protein has recently been solved it has been possible to assess the accuracy of the predictions made. The main problems with the approach seem to be that the method models TM helices as ideal helices arranged in parallel and the suggestion that the protein was a dimer. Large deviations from these assumptions, such as the kinked, highly tilted or partially-spanning helices seen in the actual structure, cause inaccuracies in the modelling.

As a first attempt at *ab initio* modelling, the current method has a number of limitations. For example, the method is unlikely to be able to select a single model with confidence for proteins that have kinked or tilted helices. However it remains useful for any TM protein: if one model scores significantly more highly than the others this implies that (i) the protein structure will be similar to that of the model and (ii) the helices of the protein are relatively straight and parallel. On the other hand, for proteins for which all models score more similarly, a single model cannot be selected in the absence of constraining mutagenesis or other experimental data. This implies an irregular helix packing arrangement, with a considerable number of tilted TM helices. In this case, the predicted buried residues on each helix will provide useful structural information and may be able to guide experimental studies to help further limit the number of possible models.

In the present work a protein showing pseudo-3-fold symmetry was studied, in order to constrain the number of possible models. However, for families for which there are no simplifying symmetry constraints, the method remains useful, provided sufficient computational resources are available to deal with the large number of potential models that

must be considered. There remain large standard deviations associated with the values of the LA scale, and more TM protein structures are needed to improve the accuracy of predictions using this scale. Finally, reliable information about the oligomeric state of the protein is essential to translating the predictions of buried helix faces into possible arrangements of TM helices for scoring. Now that these limitations have been identified, improved modelling methods can be developed which incorporate these factors.

In conclusion, this work has involved in investigation into not only UCP structure, but also into our current understanding of membrane protein structure and our ability to model it. The work has shown that, while the model proposed for the UCPs showed similarities with the correct structure, the technique is too simple and not yet good enough to predict models with confidence for proteins with tilted or kinked helices. More membrane protein structures will need to be solved, and our understanding of the ways in which they pack to obtain stability will need to increase, before improved accuracy is possible. In addition, it seems that more complex modelling procedures, in which helix tilting and kinking are considered, will be needed to increase predictive accuracy for many protein families. Therefore, although the structures of membrane proteins are limited by their unusual environment and they are less diverse than water-soluble proteins, the challenge of predicting their tertiary structure from sequence has yet to be solved.

Chapter 5

Mechanisms by which DAF-16 regulates ageing in *Caenorhabditis elegans*

5.1 Introduction

5.1.1 The insulin/IGF signalling pathway and control of lifespan

As described in the Introduction (Chapter 1), the insulin/IGF-like signalling (ILS) pathway regulates lifespan. The pathway (illustrated schematically in Figure 1.2) is highly conserved in insects, such as *Drosophila melanogaster*, nematodes, like *Caenorhabditis elegans* and mammals, such as *Mus musculus*. Despite this conservation, several regions have undergone expansion in one or more species. For example, while *Drosophila* has 7 insulin-like ligands (Brogiolo *et al.*, 2001), *C. elegans* is thought to have 37 (Pierce *et al.*, 2001). When ligands bind to the DAF-2 receptor, the signal is transferred, via the various signalling proteins shown in Figure 1.2, to a transcription factor known as DAF-16 (in C. elegans) or FOXO (in *Drosophila*). As illustrated in Figure 5.1, the phosphorylation of DAF-16/FOXO leads to it being excluded from the nucleus, both directly and indirectly affecting transcription of various target genes. DAF-16 belongs to the forkhead family of TFs. Members of the this family share a common DNA binding motif known as a forkhead, or winged helix, domain (Greenberg & Boozer, 2000) and appear to have important roles in controlling development (Hope *et al.*, 2003).

The many functions of the ILS pathway are complex. In fact, rather than a simple linear pathway, it is perhaps more appropriate to consider the pathway as a network, with multiple inputs that are integrated and processed before conversion to multiple outputs. One method for attempting to establish the processes controlled by this pathway is to

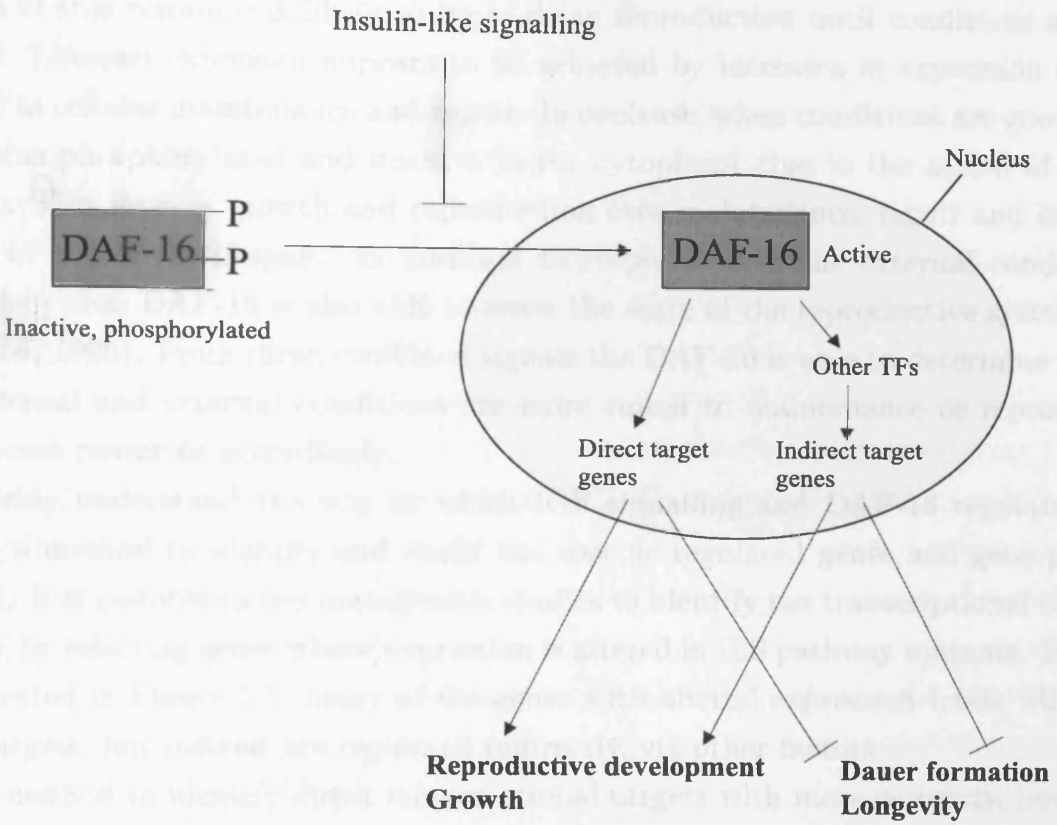


Figure 5.1: A simplified view the mechanism by which ILS inactivates DAF-16 and controls lifespan, illustrating the concept of direct and indirect DAF-16 target genes. TFs: transcription factors.

ablate various members of the cascade and identify phenotypic changes in the resulting mutants. This method has suggested a role for the insulin/IGF pathway in the control of both growth and development (Gems *et al.*, 1998; Brogiolo *et al.*, 2001) and lifespan (Kimura *et al.*, 1997a; Tatar *et al.*, 2001; Clancy *et al.*, 2001; Holzenberger *et al.*, 2003).

The role of ILS signalling in ageing appears to be evolutionarily conserved from nematodes (Ogg *et al.*, 1997) to mammals (Holzenberger *et al.*, 2003; Bluher *et al.*, 2003). Similarly, the pathway regulates DAF-16 transcription factors in a range of organisms. DAF-16 has been shown to activate longevity-promoting genes and repress genes that accelerate ageing (Lee *et al.*, 2003; Ookuma *et al.*, 2003; Murphy *et al.*, 2003). It therefore seems likely that the targets of DAF-16 are the direct biochemical determinants of ageing in many species, possibly including humans.

As described in Chapter 1, the current theories suggest that DAF-16 acts as a switch to divert resources between growth and reproduction or cellular maintenance and repair (Henderson & Johnson, 2001). When food is scarce or conditions are otherwise poor, DAF-

16 enters the nucleus, leading to an extension of lifespan and a reduction in fecundity. The function of this response is likely to be to delay reproduction until conditions are more suitable. Lifespan extension appears to be achieved by increases in expression of genes involved in cellular maintenance and repair. In contrast, when conditions are good, DAF-16 remains phosphorylated and inactive in the cytoplasm, due to the action of the ILS pathway. This favours growth and reproduction over maintenance, repair and longevity, leading to a normal lifespan. In addition to responding to the external conditions it seems likely that DAF-16 is also able to sense the state of the reproductive system (Hsin & Kenyon, 1999). From these combined signals the DAF-16 is able to determine whether both internal and external conditions are more suited to maintenance or reproduction, and allocate resources accordingly.

To truly understand the way in which IGF signalling and DAF-16 regulate ageing requires a method to identify and study the specific regulated genes and gene products involved. It is possible to use mutagenesis studies to identify the transcriptional targets of DAF-16, by selecting genes whose expression is altered in ILS pathway mutants. However, as illustrated in Figure 5.1, many of the genes with altered expression levels will not be direct targets, but instead are regulated indirectly, via other factors.

One method to identify direct transcriptional targets with more certainty, however, is via the analysis of the non-coding sequences surrounding the genes that are differentially expressed in ILS mutants. It is possible to look for correlations between the occurrence of a particular transcription factor binding site in these regions and the expression of the associated genes, using techniques described in Section 5.1.2.1. Here, DAF-2 mutants will be compared against DAF-2/DAF-16 mutants, to identify patterns of gene expression present when DAF-16 is constitutively on or off respectively. This may allow DAF-16 binding elements to be identified in direct targets, and other elements to be found in other longevity-associated genes controlled by different transcription factors.

Several previous studies have made the current work possible. Firstly, Kim *et al.* (2001) produced a gene expression 'topomountain' map, by clustering *C. elegans* genes that showed similar expression patterns across multiple studies and a range of conditions. The authors found that some 'mountains' were enriched for genes expressed in the same cell types, where as others contained genes functionally related by being involved in the same cellular process, such as heat shock proteins or collagens. These mountains enable predictions to be made about the role of unknown genes, according to the characteristics of the mountain to which they belong. The study has also provided large groups of co-regulated genes, whose expression can be compared between wild-type and long-lived mutants to identify general patterns.

Such a comparison was performed by McElwee *et al.* (2004), who annotated all *C.*

elegans genes, using the topomountain groups produced by Kim *et al.* (2001), and gene families from INTERPRO (Apweiler *et al.*, 2000) and GO (Ashburner *et al.*, 2000). They then compared expression between DAF-2/DAF-16 and DAF-2 mutants using EASE (Hosack *et al.*, 2003). EASE detects significant over-representation of particular classes of genes within a set of co-regulated genes, compared to the genome as a whole. The method allowed McElwee *et al.* (2004) to identify functional classes of genes which are over-represented amongst up-regulated (longevity-associated) and down-regulated (ageing-associated) genes. The 20 classes of genes most over-represented within longevity- and ageing-associated genes are shown in Table 5.2 in Section 5.2.4.

The current study analyses the transcription factor binding sites associated with each of these ageing- and longevity-associated gene groups, in order to determine by which transcription factors their expression is regulated. The value of this approach lies in the fact that the groups contain co-regulated and often functionally related genes, permitting different transcriptional control elements to be identified that are specific for each group. In contrast, the combined analysis of all DAF-16 regulated genes would be unlikely to uncover many significant patterns of motif over-representation, due to the great variability of genes, and therefore of transcriptional mechanisms, included. This work has allowed the genes to be classified into direct and indirect targets of DAF-16 and other transcription factors involved in control of longevity to be identified. These results are discussed with respect to the role each transcription factor or class of genes may play in the regulation of longevity by DAF-16. As a result, it is hoped that our understanding of the role of DAF-16 in longevity, and the mechanisms of ageing itself, will be increased.

5.1.2 Transcription factors

Transcription factors (TFs) are proteins that regulate the expression of genes, and hence the synthesis of proteins. In this way they can control the function and development of the cell, and consequently the whole organism. They function by binding to specific transcription factor binding sites in the DNA, generally upstream of the gene they control, in the promoter region. Transcription factors are modular proteins consisting of several discrete domains, each with a defined function. These will include a DNA-binding domain, such as the forkhead domain of DAF-16, and one or more activation or repression domains which interact with the DNA or with other transcription factors or cofactors to affect gene expression.

Each transcription factor controls the expression of a specific gene or set of genes, via a characteristic binding site. Generally, transcription factor binding sites (TFBSs) are less than 10bp in length. They are often degenerate (so that more than one base can

be tolerated at certain positions). Transcription factors are often compartmentalised to particular cell types. Some are continually present and others are released or activated in response to certain signals, either from the external environment or from within the cell itself. They therefore function to integrate environmental signals and convert them into the appropriate response within the cell. Identification and characterisation of TFBSs is important for understanding how this process occurs, and can provide clues about the control of complex biological processes.

TFBS motifs are often represented as weight matrices, rather than simple consensus sequences. These can be obtained using a variety of experimental techniques such as nuclease or restriction enzyme footprinting (Hardenbol *et al.*, 1997; Wilson *et al.*, 2001; Papavassiliou, 2001), binding site selection (Baes & Declercq, 1998) or methylation protection (Shaw & Stewart, 1994; Reid & Nelson, 2001). Such methods are used to identify a set of sequences to which a particular TF will bind. These sequences are then aligned and converted to a matrix, in which at each position the probability of observing each of the 4 bases is described. As a result, each occurrence of the motif observed is considered as just one example of all possible sequences the motif can take, with a probability attached to it. The importance of the use of these matrices is that they are able to increase the accuracy of TFBS identification by permitting degeneracy and distinguishing between mismatches of different severity (Staden, 1984; Quandt *et al.*, 1995).

In order to achieve integrated control of the cell, each gene is likely to be controlled by multiple, interacting factors and contain multiple binding sites within its promoter (Yuh *et al.*, 1998). There is therefore a need to identify and characterise as many transcription factors and their binding sites as possible and to give consideration to the interactions between these proteins when investigating transcriptional control. An understanding of these processes is crucial to understanding the gene regulatory networks that define cell function. Some of the computational techniques that can be used are described in the following section.

5.1.2.1 Methods for the identification of transcription factor binding sites

There are two main computational methods that have been used to identify transcription factor binding sites. These are often termed the 'single species, multiple gene' and the 'single gene, multiple species' approaches. Examples of both of these methods are reviewed in more detail by Duret & Bucher (1997). More recently, methods which combine the two techniques have been developed in an attempt to make use of the strengths, whilst avoiding the disadvantages, of each. Several experimental techniques have also been used, although the present large-scale sequencing initiatives have fuelled the need for more

rapid, computational approaches in addition. However, it should be stressed that all computational methods for transcription factor binding site identification only identify *potential* functional TFBSs. Experimental confirmation of the identified candidate motifs is therefore necessary.

Single species, multiple gene approaches Single species, multiple gene approaches generally make use of motif searching algorithms. These algorithms are used to search for motifs that are over-represented in the promoter sequences of otherwise unrelated genes, often that have been shown to be co-regulated.

The single species, multiple gene approach can be divided into knowledge-based and blind search methods. Knowledge-based methods search for previously identified motifs, often using libraries of weight matrices (Quandt *et al.*, 1995; Chen *et al.*, 1995; Prestridge, 1996). Pre-computing and verifying the weight matrices in this way enables searches for TFBSs to be completed much more accurately, but no more slowly, than methods using simple consensus sequences (Quandt *et al.*, 1995). Examples of this class of approach are Match (Kel *et al.*, 2003) and Clover (Frith *et al.*, 2004). These methods search promoters for instances of known TFBSs, such as those that are stored in the Transfac Database (Wingender *et al.*, 2001).

In contrast to the knowledge-based methods, blind-search techniques align the promoters, usually of a set of co-regulated genes, to identify over-represented motifs that are candidates for TFBSs. These methods tend to be slower and more computationally expensive than searching for known motifs but they are necessary to identify new motifs that have not previously been characterised. Examples of blind-search methods are AlignACE (Hughes *et al.*, 2000) and MEME (Bailey & Elkan, 1994). These methods output alignments of over-represented motifs and an associated E-value. One difficulty with these approaches is that there is no simple method to compare the identified motif with those already known to determine whether it is novel or whether its binding factor is known. Often, due to the lack of completeness of the databases, the only method is to compare the motifs to the literature to assign them to known TFs.

Single gene, multiple species approaches Single gene, multiple species approaches are often referred to as phylogenetic footprinting (Tagle *et al.*, 1988). There are two main varieties of phylogenetic footprinting. The first method requires the alignment of promoter regions from orthologous genes from a range of species. Algorithms are then used to identify regions that are more conserved than the surrounding sequence. TFBSs are assumed to be more conserved than adjacent regions since selective pressure to maintain functional sites will tend to lead to slower evolution. Many studies have

shown the validity of this approach by identifying TFBSs that have previously been experimentally determined (Blanchette & Tompa, 2002; Cliften *et al.*, 2003; Berezikov *et al.*, 2004). However, the results of phylogenetic footprinting studies are very sensitive to the alignment program used and the divergence of the species analysed (Cliften *et al.*, 2003; Berezikov *et al.*, 2004).

In contrast to the alignment-based methods described above, the algorithm Footprinter uses the species tree as an approximation of the phylogenetic relationships between the orthologous sequences (Blanchette & Tompa, 2002; Blanchette *et al.*, 2002; Blanchette & Tompa, 2003). It then identifies motifs for which the parsimony score (the number of mutational changes) is unexpectedly low, given the divergence of the species in which the motif was found. The accuracy of the approach is improved by using relatively diverged sequences to minimise background conservation but reducing the effects of misalignment and of binding site turnover by only requiring that motifs are found in a subset of the species. Also, because it searches not only for TFBSs that are identical between species, but also those with a low parsimony score, its false negative rate is further reduced.

The phylogenetic footprinting method assumes that regulatory mechanisms are conserved between species. Whilst this is generally thought to be the case, the regulatory motifs of several gene families have been missed by this method due to divergence that has occurred (Blanchette & Tompa, 2002). It has also been suggested that there is considerable turnover of TFBSs, so that 30-40% of human TFBSs are not functional in rodents (Dermitzakis & Clark, 2002). Regulatory elements that have been fixed relatively recently in evolution may also be missed.

Multiple gene, multiple species approaches More recently there has been a move to try to combine the two approaches described above, leading to development of a 'multiple gene, multiple species' approach. One particular method, named PhyloCon, locally aligns the promoters of orthologous genes and uses these alignments to generate profiles of each promoter (Wang & Stormo, 2003). These profiles are then compared between co-regulated genes, in order to identify conserved motifs. Improvements in performance over other techniques were identified using this method. Regression-based techniques, such as MOTIF REGRESSOR (Conlon *et al.*, 2003), have also been successful. In addition, various improvements have been suggested, such as the explicit use of phylogenetic trees or permitting gaps in motifs, that are likely to increase accuracy in the future.

5.1.3 Aims

The current chapter aims to exploit available TFBS analysis tools and microarray data concerning gene expression levels during ablation of DAF-2 and DAF-2/DAF-16, in an attempt to:

1. identify potential DAF-16 binding sites
2. identify genes that are potential direct transcriptional targets of DAF-16
3. investigate functional classes of DAF-16 regulated genes with a potential evolutionary conserved role in ageing
4. identify and investigate the factors controlling the expression of other potential longevity determining genes (indirect targets of DAF-16)
5. identify potential transcription factors or hormones within the direct DAF-16 targets that may contribute to downstream effects on lifespan
6. ultimately to increase our understanding of the mechanisms by which DAF-16 controls lifespan

It is hypothesised that DAF-16 initiates a regulatory cascade, similar to that illustrated in Figure 5.2, that leads to coordinate control of a large number of genes that determine lifespan. Some longevity-determining genes will be directly bound by DAF-16 itself. In addition, some of the targets of DAF-16 are hypothesised to be TFs or hormones, which in turn may initiate expression or repression of a further set of indirect target genes, in response to DAF-16 activity. Throughout this chapter it is hoped that these proteins, and the direct and indirect targets of DAF-16, will be identified.

At the end of the chapter, in Section 5.4, we will return to these aims and discuss how fully they have been achieved.

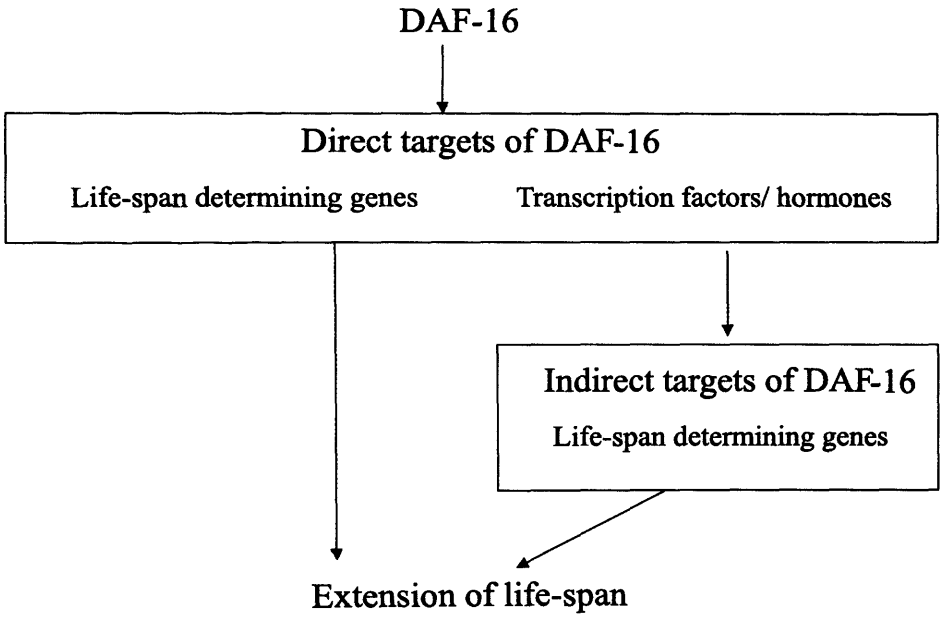


Figure 5.2: Hypothesised mechanism by which DAF-16 regulates lifespan.

5.2 Methods

5.2.1 Overview of methods

The methods used throughout this chapter attempt to divide and sub-divide the longevity- and ageing-associated gene classes identified by McElwee *et al.* (2004) according to their transcriptional control. Firstly, as shown in Figure 5.3, longevity-associated gene classes were divided into direct and indirect targets of DAF-16. Then, as far as possible, within the direct and indirect targets, sets of gene classes controlled by different combinations of other TFs were identified. This will permit a clearer understanding of the mechanisms by which lifespan is controlled by the interaction of DAF-16 with other TFs.

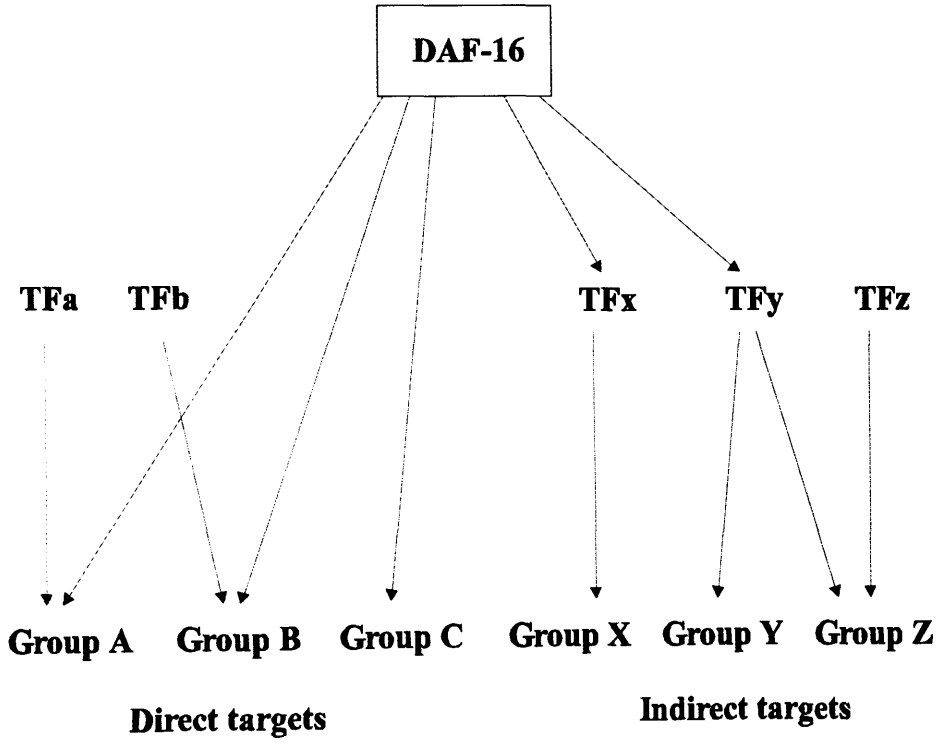


Figure 5.3: Illustration of the division of DAF-16 target gene classes into groups regulated by similar TFs.

The major tools used throughout this work are listed in Table 5.1.

Method	Address and notes
Clover Frith *et al.* (2004)	*http://zlab.bu.edu/clover* Tool for detection of TFBS over-representation
TransFac Wingender *et al.* (2001)	*http://www.gene-regulation.com/pub/databases.html#transfac* Database of TFs and DNA-binding profiles
TFBlast Wingender *et al.* (2001) Altschul *et al.* (1997)	*http://www.gene-regulation.com/cgi-bin/pub/programs/tfblast/tfblast.cgi?* BLAST-based tool to identify TFs from sequence
CisOrtho Bigelow *et al.* (2004)	*http://www.dev.wormbase.org/db/cisortho/query* Database of predicted *C. elegans* and *C. briggsae* orthologue pairs

Table 5.1: Online tools for TFBS analysis used in this work.

5.2.2 Identification of transcription factor binding sites by Clover

The genes within each of the functional classes described in Section 5.2.4 were analysed using Clover. Clover uses the matrices in Transfac to identify possible TFBSs in a set of sequences. It then calculates a raw score, indicating the predicted occupancy of the TFBS, based on a thermodynamic model. This score will be high if the sequences match very closely to the matrix. Scores for multiple sites are combined, for example if the site is repeated several times within a single promoter, or if weak 'shadow' sites exist, partially overlapping or surrounding a stronger site. Multiple copies of a site are thought to increase binding affinity and occupancy, perhaps by guiding the TF along the DNA to the correct site, or perhaps by allowing cooperative binding of multiple factors.

The raw score for each TF binding matrix is then compared between the gene set of interest and a background set of sequences from the same genome. A probability (P) value is calculated to indicate the degree of over- or under-representation of TFBS occupancy within the studied gene list, compared to the background. Multiple testing is used (1000 randomisations), in order to increase the accuracy with which the probability is calculated. Since a similar level of false positives would be expected between the gene list and background, any difference in score between the two sets is likely to be caused by functional sites. In this way, the level of false positives is much lower than if all individual genes containing hits against the binding matrix were reported as targets of that TF.

The genes classes described in Section 5.2.4 were analysed using Clover with all default parameters. The 5' and first intron sequences were analysed separately, and the background used was all 5' or first intron sequences, respectively, in the whole *C. elegans* genome. The matrices used to search each set of genes were (1) All 582 insect and vertebrate matrices in Transfac Professional, v8.1; (2) All nematode matrices and consensus sequences from Transfac Professional, v8.1 (Binding consensus sequences were used, where a matrix was unavailable, to generate an artificial matrix for searching); (3) Several ageing-associated matrices from the literature, shown in Figure 5.4.

DAF-16 BINDING ELEMENT ASSOCIATED ELEMENT (DAE)	HEAT-SHOCK ELEMENT (HSE)	HEAT-SHOCK ASSOCIATED SEQUENCE (HSAS)
	AAGCCCGG	TTGTCCT
CTTATCA	GTCTCGGA	GAGTCAT
CTTATCA	TTCTCGAA	GAGTTAT
CTTATCA	TTCTCGAA	GGGTTTT
CTTATCA	TTCTAGAA	GGGTTTC
CTTATCA	TTCTAGAA	GGGTTTC
CTTATCA	TTCTAGAA	GGGTTTC
CTTATCA	TTCTAGAA	GGGTGTC
CTTATCA	TTCTAGAA	GGGTGTC
CTTATCA	TTCTAGAA	GGGTGTC
CTTATCA	TTCTAGAA	GGGTGTC
	TTCTAGAA	GGGTGTC

DAF-16 BINDING ELEMENT (fDBE)	HYPOXIA RESPONSE ELEMENT (-LIKE) (HRE)
GTGGGTTGTTTACATCTA	CGGTACGTGATTTT
AATTGTTGTTTACTGAAC	GTCTACGTGACTTT
TTGTATTGTTTACGXXXX	AAAAACGTGACAAT
TGCTTTTGTTTACATTTT	CTCCACGTGATATC
TATTGTTGTTTACGACAT	GCGTACGTGATAAT
GTATGTTGTTTACGTCAT	CTATACGTGAGTCA
GATCTTTGTTTACCCCTA	GGATACGTGACAAC
GATGATTGTTTACTACTA	TGGAACGTGAAAAG
AGGATTTGTTTACGTGTA	GTGAACGTGATAAA
CGATCTTGTTTACAACTT	GCCTACGTGATTTA
GTGAGTTGTTTACXXXXX	TAGAACGTGAGAAA
GGACCTTGTTTACTATTT	CAGCACGTGACGCA
CGATGTTGTTTACCTTGA	AACTACGTGAAGAA
CGAGGTTGTTTACAAATT	ACCCACGTGAAGCT
CGTGATTGTTTACACAAA	GCAAACGTGAATAA
AGGGGTTGTTTACAATAT	GCGTACGTGATTAG
GAGCTTTGTTTACAATTT	GTCTACGTGAGAAA
TAGAGTTGTTTACGTACT	AAATACGTGAATGA
CGCTATTGTTTACAAATA	ATTCACGTGATCAA
AGGAGTTGTTTACAACGT	AGAAACGTGAATTT
CCTGCGTGTTTACAATTT	GATCACGTGAATGA
TTACGTTGTTTGTGAATA	CCTTACGTGATAAT
TCTTGTTGTTTTTCTCGT	CTTAACGTGATAAA
CAGCGATGTTTACCCCTT	TTAAACGTGAATTA
GGGTGTTATTTACATXXX	AAAAACGTGATAAA
XXXGATTGTTTGTTTCAG	TCTCACGTGACTCA
XXXXGATGTTTACACTTA	AAAAACGTGAAAGA
TTGTTTTGTTTATCCCCT	AGTCACGTGAGAAT
	CGCTACGTGATATT
	GAGTACGTGACATA
	CTAAACGTGAATTT
	GCAAACGTGACATA

Figure 5.4: Ageing-associated matrices from the literature that were used in this study. They are two heat-shock matrices, the heat shock element (HSE) and heat shock associated element (HSAS) (GuhaThakurta *et al.*, 2002), the DAF-16 binding-element (fDBE) (Furuyama *et al.*, 2000), the DAF-16 associated element (DAE) (Murphy *et al.*, 2003) and the Hypoxia-response element (HRE) (personal communication, J.A. Powell-Coffman).

TFBSs over- or under-represented with a P value of less than 0.01 were reported and the motifs identified were compared between the different classes of genes. Over- or under-representation of TFBSs corresponding to different TFs from the same sub-family, according to Transfac annotations or the published literature, were combined. This gave a non-redundant list of binding sites for each gene class analysed.

The longevity-associated gene classes were classified as direct targets of DAF-16 if they were significantly enriched for the DAF-16 binding element (DBE) from Transfac (N$DAF16_01) ($P < 0.01$). In contrast, indirect targets of DAF-16 were those genes whose expression increased in DAF-2 mutants but that lacked over-representation of the DBE. The gene classes were grouped into 'regulatory sets' which contained similar combinations of over-represented regulatory elements. The literature was searched for information regarding important TFs from each regulatory set, in order to attempt to define a mechanism by which each may contribute to a long-lived phenotype.

5.2.3 Identification of transcription factors whose expression is altered in DAF-2 mutants

This was performed using TFBlast, a BLAST-based tool (Section 1.3.1) available as part of the Transfac website. Coding sequences of all proteins with altered mRNA expression in DAF-2 mutants were downloaded from EnSmart (Kasprzyk *et al.*, 2004) and BLASTED against the sequences of all known TFs found in Transfac. Matches with an E-value of less than 1^{-20} and with expression fold changes of greater than 1.5 or less than 0.5 were selected for further examination. This further examination involved running PSI-BLAST (Section 1.3.1) (for a maximum of 20 iterations or until convergence, using a threshold of $1e^{-40}$) to identify sequence relatives. (In the case of transcript F26D12.1, this stringent threshold gave only one sequence relative, so a value of $1e^{-20}$ was used). The sequences of these relatives and of the query sequence were aligned using CLUSTALW (Thompson *et al.*, 1994) and the resulting phylogenetic tree was generated using PHYLIP (Felsenstein, 1993). Those query sequences whose closest identified orthologue was a known TF were classified as likely TFs themselves. Possible *C. elegans* homologues were identified from WormBase (WormBase web site, *http://www.wormbase.org*, release WS120, March 2004) and the literature. The function of these TFs, and their known targets, are discussed with respect to a possible role of these TFs in the control of lifespan.

5.2.4 Dataset of analysed genes and sequence collection

The sets of genes analysed were those associated with various functions implicated in the control of lifespan by McElwee *et al.* (2004). These authors used Affymetrix whole genome oligonucleotide microarrays to measure gene expression levels in various mutants of *C. elegans*: (i) *daf-2(m577)*; (ii) *daf-2(m577), DAF-16*; (iii) *daf-2(e1370)* and (iv) *daf-2(e1370), DAF-16*. Since DAF-2 inactivates DAF-16, this method enabled a comparison of gene expression between mutants where DAF-16 activity was constitutively on (i and iii) or off (ii and iv). Two *daf-2* mutant alleles were used to reduce the likelihood of detecting allele-specific changes in gene expression that are unrelated to ageing. Statistical methods (see McElwee *et al.* (2004)) were used to identify a list of 1348 up-regulated and 926 down-regulated genes, with a median false discovery rate of 5%.

McElwee *et al.* (2004) annotated these lists of differentially-expressed genes, using information from a number of databases and the expression topomountain map of Kim *et al.* (2001) (Section 5.1.1), and analysed them using EASE. EASE is described as a tool for automatically converting a list of genes, derived from studies such as microarray analysis, to a set of functional 'themes' associated with that list (Hosack *et al.*, 2003). This is achieved by detecting functional classes of genes that are significantly over-represented amongst the gene list of interest. McElwee *et al.* (2004) identified a number of gene classes over-represented amongst the up-regulated and down-regulated genes in DAF-2 mutants. These gene classes are likely to be associated with longevity and ageing respectively.

The 20 functional classes most strongly associated with longevity and 20 classes most strongly associated with ageing, according to the McElwee *et al.* (2004) study, were selected for analysis. These functional classes are shown in Table 5.2. The value of dividing the longevity- and ageing-associated genes into groups according to their function lies in the ability to investigate the different transcriptional control of each of the classes of genes regulated by DAF-16.

Class	Function	Number of genes	EASE score
GO0004497	Monooxygenase activity	22	1^{-8}
GO0016491	Oxidoreductases	35	1^{-6}
	EASE ageing-associated classes: down-regulated		
Class 2 genes	Potential ageing-associated genes, defined by Murphy *et al.* (2003)	59	1^{-17}
Two-fold down	Genes down-regulated by two-fold or more as animals recover from dauer (Wang & Kim, 2003)	141	1^{-13}
Four-fold down	Genes down-regulated by four-fold or more as animals recover from dauer (Wang & Kim, 2003)	71	1^{-6}
Mount 8	Intestine, antibacterial, UGTs	81	1^{-15}
Mount 19	Amino acid and lipid metabolism, cytochrome P450	71	1^{-50}
Mount 21	Lipid metabolism	33	1^{-15}
Mount 24	Amino acid and lipid metabolism, fatty acid oxidation	49	1^{-32}
Mount 27	Amino acid metabolism, energy generation	28	1^{-18}
Mount 31	Unknown function	10	1^{-7}
Lipid metabolism	Lipid metabolism, defined by Wang & Kim (2003)	43	1^{-9}
IPR002213	UDP-glucuronosyl/UDP-glucosyltransferases	18	1^{-8}
UGT	UDP-glucuronosyl/UDP-glucosyltransferases, defined by Wang & Kim (2003)	18	1^{-8}
Transporters	General substrate transporters, defined by Wang & Kim (2003)	47	1^{-7}
GO0016758	Transferases, transfer hexosyl groups	17	1^{-7}
IPR005828	General substrate transporter, transport small solutes in response to chemiosmotic ion gradients	17	1^{-6}
GO0006810	Transport proteins, involved in the transport of all substrates	47	1^{-5}
IPR003366	Protein of unknown function DUF141	29	1^{-23}

Table 5.2: *continued on next page*

Class	Function	Number of genes	EASE score
IPR004119	Protein of unknown function DUF227	9	1^{-5}
IPR005071	Protein of unknown function DUF274	11	1^{-9}
IPR001304	C-lectins	26	1^{-5}

Table 5.2: Ageing- and longevity-associated functional classes of genes analysed. See Kim *et al.* (2001) for a description of the methods used to generated the 'Mount' groups. EASE scores are taken from McElwee *et al.* (2004). The lower the score the greater the over-representation of a functional class with ageing- or longevity-associated genes.

Sequences used in the analysis were taken from the e_snip file, downloaded from CisOrtho (Bigelow *et al.*, 2004). The sequences of the 5' upstream region (including the 5' untranslated region) and of the first intron were extracted from this file for (1) all genes in the *C. elegans* genome (2) all genes for that are up- or down-regulated by DAF-16 and classified as belonging to each ageing-associated functional group described in Table 5.2. The length of 5' sequences were limited to 1000bp, since few TFBSs are found more than this distance from the transcription start site.

The Ensembl tool EnsMart (Kasprzyk *et al.*, 2004) was used to obtain the protein coding sequences of the genes for which expression had been determined, for use with TFBlast (see Section 5.2.3).

5.3 Results

5.3.1 Longevity-associated genes and their regulation

5.3.1.1 Identification of direct and indirect DAF-16 targets

In the introduction to this chapter it was hypothesised that DAF-16 initiates a cascade, shown in Figure 5.2, that regulates lifespan. This work has enabled us to add considerable data to this hypothesised cascade. In particular, likely direct and indirect DAF-16 target gene classes have been identified, as summarised in Table 5.3 and Figure 5.5. Strong over-representation of DAF-16 binding elements (DBEs) are observed in all direct target gene classes ($P<0.01$ in all cases).

Direct targets	Indirect targets
Mount 8 (intestine, antibacterial, UGTs)	Cytochrome P450s (IPR002401/2403/1128)
Mount 15 (unknown function)	Monooxygenases (GO0004497)
Mount 17 (collagen)	Glutathione-S-transferases (IPR004045/4046)
Metabolism (GO0008152)	Transposases (IPR001888)
Oxidoreductases (GO0016491)	Mount 6 (neuronal genes)
Dauer-specific tag genes	

Table 5.3: Summary of direct and indirect DAF-16 target gene classes, as identified during this work.

5.3.1.2 The role of longevity-associated gene groups in the control of lifespan by DAF-16

A wide range of gene classes are implicated in the response to DAF-16. While the cytochrome P450s, UDP-glucuronyl transferases and glutathione-S-transferases have previously been proposed to have a role in the control of longevity by DAF-16 (Murphy *et al.*, 2003; McElwee *et al.*, 2004), the mechanisms by which this occurs require considerable further investigation. This section discusses with respect to the literature what is currently thought to be the role of each of the longevity-associated gene groups in the control of lifespan.

Metabolic genes The role of metabolic genes in the control of longevity is difficult to interpret mechanistically, due to the extremely wide variation in function between the genes in this class (GO0008152, General metabolism). However, many of these genes appear to be direct targets of DAF-16, and may therefore play a role in longevity determination.

Theory suggests that lifespan is extended when metabolic resources are diverted from growth and reproduction to maintenance and repair, and that this balance may be controlled by DAF-16 (Henderson & Johnson, 2001). Hence we would expect the up-regulated metabolic genes to be specifically involved in the latter processes. The up-regulation of general metabolic genes is not related to lipid or protein metabolism, since as shown in Table 5.2, many gene classes related to these functions are strongly down-regulated in DAF-2 mutants. Instead, up-regulation of metabolic genes is likely to be due to the up-regulation of a number of other functional classes of genes within GO0008152. Possible candidates, highly consistent with the current theories of ageing, are genes involved in the metabolism of drugs, toxins, xenobiotics, hormones and reactive oxygen species and genes involved in the regulation of metabolism itself.

Mount 6 Mount 6 is enriched in neuronal genes. These genes in general appear to be indirect targets of DAF-16 associated with longevity. The finding that neuronal genes are involved in the control of lifespan, is consistent with the work of Wolkow *et al.* (2000), who have shown the importance of neuronal gene expression in lifespan extension. The authors showed that expression of DAF-2 specifically in neurones, but not in muscle or intestine, is sufficient to restore normal longevity of knockout animals. They concluded that DAF-16 causes changes in neuronal gene expression, leading to the release of a cell non-autonomous factor, that is proposed to regulate lifespan. It follows that this set of neuronal genes is likely to contain key longevity-determining genes, and experimental analysis of this set is likely to be fruitful. However, as discussed later, more recent work has failed to confirm the importance of neuronal genes in control of lifespan (Hsin & Kenyon, 1999; Libina *et al.*, 2003). It will be important to determine whether the DAF-16 regulated Mount 6 genes are indeed neuronal. The identity of these genes requires further investigation.

Mount 17 Despite the enrichment of Mount 17 for collagens, only 3 of the 40 genes in this group that are up-regulated in DAF-2 mutants have this function (McElwee *et al.*, 2004). Hence the over-representation of Mount 17 genes amongst up-regulated genes is likely to be due to the role of other, non-collagen, genes in longevity.

Genes involved in xenobiotic metabolism It has been proposed that one mechanism by which lifespan is extended may be via the up-regulation of genes involved in detoxification (McElwee *et al.*, 2004; Murphy *et al.*, 2003). Such genes may function by protecting the cell against xenobiotic and other toxic compounds, which may contribute to the damage that accumulates with age. Genes involved in xenobiotic detoxification include:

- cytochrome P450s (CYPs). These enzymes are non-specific monooxygenases, of great clinical importance (Werck-Reichhart & Feyereisen, 2000), that act on a wide range of compounds leading to metabolism and synthesis of hormones and activation and detoxification of drugs (reviewed in Omiecinski *et al.* (1999)).
- UDP-glucuronyl transferases (UGTs). The UGTs are found in the endoplasmic reticulum, where they glucuronidate small lipophilic molecules, solublising them to permit excretion.
- Glutathione-S-transferases (GSTs). GSTs detoxify electrophilic compounds either by adding the tripeptide glutathione, via peroxidase activity or by passive binding.

Several classes of xenobiotic detoxification genes are up-regulated in long-lived animals (McElwee *et al.*, 2004; Murphy *et al.*, 2003) and in dauers (McElwee *et al.*, 2004; Wang & Kim, 2003). Murphy *et al.* (2003) have directly demonstrated a role for CYPs and UGTs in longevity using RNAi. Hence there is a growing body of evidence in support of the role of xenobiotic metabolism in longevity. However, the mechanisms by which these genes are activated remain poorly understood.

The current study suggests that antibiotic proteins and the UGTs (Mount 8) are likely to be direct targets, while CYPs and GSTs appear to indirect targets of DAF-16. As shown in Table 5.5, GSTs appear to be regulated by heat shock factors, consistent with the proposal, described in Section 5.3.1.3, that both DAF-16 and heat shock proteins are required for extension of lifespan (Hsu *et al.*, 2003). It can therefore be hypothesised that, in unfavourable conditions, both DAF-16 and heat shock factors are activated, which in combination stimulate the expression of CYPs, GSTs and UGTs, leading to extension of lifespan.

Intestinal genes In addition to antibacterial and UGT genes, Mount 8 is enriched with intestinal genes. These genes are up-regulated in long-lived, DAF-2 mutants and appear to be direct transcriptional targets of DAF-16. In *C. elegans*, the intestine has a lipid-storage role, similar to that of adipose tissue in mammals. An important role for the intestine in nematode ageing has been shown by Libina *et al.* (2003). These authors have shown that expression of DAF-16 in the intestine alone is sufficient to cause a 50% extension of lifespan in DAF-2 mutants, although expression in other tissues is also required for full extension. Expression of DAF-16 in the intestine fully restores wildtype lifespan in germ-line deficient animals, suggesting that the germ-line signals that control lifespan Hsin & Kenyon (1999) are mediated through the intestine.

In addition, insulin-like peptides are expressed in the intestine and found in Mount 8 (Kim *et al.*, 2001) and their expression is regulated by the DAF-2 pathway (Murphy *et al.*, 2003). Libina *et al.* (2003) have suggested that the intestine may act as the pancreas of *C. elegans*, secreting insulin-like peptides in response to food, providing feedback to DAF-16 about the level of nutrient supply, and perhaps also contributing to the cell non-autonomous effects of DAF-16 (Apfeld & Kenyon, 1998). However, as discussed in Section 5.3.4, the insulin-like peptides studied in the current work are all believed to be expressed primarily in the nervous system (according to WormBase annotations).

Another function may contribute to the up-regulation of Mount 8 intestinal genes in long-lived mutants. Many of the Mount 8 intestinal genes are likely to be involved in protection against bacterial infection, including enzymes that degrade bacterial cell walls, proteins and DNA (Kim *et al.*, 2001). Some Mount 8 genes, such as metallothionein,

are involved in the binding and detoxification of heavy metals and other toxins. Hence Mount 8 genes may extend lifespan by protection of the organism from damage induced by toxins and bacterial infections. The importance of the intestine in control of lifespan may therefore be linked to its relatively high exposure to toxins and infection. In support of this hypothesis, it has recently been shown that long-lived ILS pathway mutants are more resistant to bacterial infection than wildtype (Laws *et al.*, 2004). In summary, it seems that the intestine plays an important role in control of lifespan in *C. elegans.*

Mount 15 Mount 15 is by far the most over-represented group in the longevity-associated genes (McElwee *et al.*, 2004). It contains a large number of genes of unknown function. That these genes appear to be direct targets of DAF-16 suggests that they may be important determinants of longevity and they therefore require further study. In addition, all Mount 15 genes are also up-regulated in dauers, suggesting a link between these genes and the longevity of dauers and DAF-2 mutants (McElwee *et al.*, 2004). Heat shock factor-1 (HSF-1) is another TF important in longevity that is found in Mount 15. While it is not up-regulated in DAF-2 mutants this is probably due to the fact that its action is controlled post-translationally (DiDomenico *et al.*, 1982).

While there is some evidence for the importance of Mount 15 in ageing, McElwee *et al.* (2004) have shown that Mount 15 is the least evolutionarily conserved of all of the longevity-associated gene groups. This suggests that these genes may be less important in the control of lifespan in other species than in *C. elegans*, where they may have undergone expansion linked to dauer formation.

Transposases Both transposases in general (IPR0001888) and the mariner transposase subclass were up-regulated in long-lived DAF-2 mutants. Transposases are enzymes which excise and insert mobile genetic elements within DNA. As described by McElwee *et al.* (2004), up-regulation of transposase activity would not be expected to lead to extension of lifespan, and is more likely to be a consequence of other DAF-2 regulated changes in gene expression or chromatin structure.

5.3.1.3 Other gene groups believed to have a role in the control of lifespan by DAF-16

While the heat shock, antioxidant and UGT gene classes were not included amongst the 20 analysed longevity-associated gene classes taken from the McElwee *et al.* (2004) study, they are thought to play a role in the control of lifespan (Honda & Honda, 1999; Murphy *et al.*, 2003; Hsu *et al.*, 2003; Lee *et al.*, 2003; Walker & Lithgow, 2003; Li *et al.*, 2004a; McElwee *et al.*, 2004). The expression of individual genes in each of these groups was

therefore investigated, as shown in Table 5.4. While the whole gene classes were not strongly up-regulated in DAF-2 mutants, certain individual members of each class were up-regulated. Many of these up-regulated genes are potential direct targets of DAF-16. It should therefore be remembered that not all heat shock proteins, antioxidant enzymes and UGTs show a similar association with longevity, and that these proteins do not necessarily play a key role. Instead, as described in the following sections, it seems that while some members of each family promote longevity, others promote ageing and that many other groups of genes play a more significant role in the control of lifespan.

It should be noted that, while DBEs were observed upstream of the potential DAF-16 direct targets in Table 5.4, there is no evidence that these motifs are functional. It will be necessary to confirm the conservation of these sites in other species (phylogenetic footprinting) to be confident that these are true direct DAF-16 targets.

Heat shock genes Heat shock proteins are chaperones and proteases that confer resistance to heat and oxidative stress by binding or degrading oxidised or denatured proteins that may cause cellular damage. In *C. elegans* the expression of heat shock proteins is controlled by the heat shock factor HSF-1. Over-expression of HSF-1 extends lifespan in *C. elegans* and HSF-1 is required for the extended lifespan of DAF-2 mutants (Hsu *et al.*, 2003). This suggests a crucial role for these proteins in the control of longevity, in which HSF-1 and DAF-16 together activate expression of longevity genes.

Given this key role, it is somewhat surprising that heat shock genes are only weakly over-represented amongst longevity-associated genes according to EASE analysis (McElwee *et al.*, 2004). (The only heat shock related group significantly over-represented in DAF-16 up-regulated genes was the Hsp20s (IPR002068) with P=0.05. All classes of heat shock protein combined showed no significant over-representation (P=0.22)). Other groups have observed up-regulation of some heat shock proteins in ILS mutants (Hsu *et al.*, 2003; Lee *et al.*, 2003; Murphy *et al.*, 2003; Walker & Lithgow, 2003; Li *et al.*, 2004a). In addition, five specific heat shock proteins are up-regulated in this study, four of which appear to be direct targets of DAF-16 (Table 5.4). Several other heat shock proteins were down-regulated in DAF-2 mutants. These results suggest that only a small number of specific heat shock proteins are involved in control of lifespan by DAF-16 and that others may play no role or even promote ageing. The molecular mechanisms behind this remain to be determined but are thought to involve repair of damaged proteins.

Both HSF-1 and DAF-16 can translocate to the nucleus and stimulate expression of certain target genes even when the action of the other is blocked with RNAi. However, these changes in expression are insufficient to cause extension of lifespan, for which both proteins are required (Hsu *et al.*, 2003). This suggests that the major longevity genes will

Protein	WormBase identifier	Fold change
Heat shock proteins		
α-B crystallin	**F38E11.1**	**12.2**
Hsp70	C12C8.1	3.7
Hsp16.1 family member	**F08H9.4**	**2.8**
Hsp20	**F43D9.4**	**2.2**
Hsp16.1 family member	**F08H9.3**	**1.7**
Members of the UGT family		
	F10D2.11	**3.6**
	C10H11.5	3.5
	T19H12.10	**2.9**
	C23G10.6	**2.5**
	T07C5.1	1.8
	H23N18.2	1.5
	F10C2.5	1.5
	F56B3.7	**1.4**
	F09G2.6	**1.2**
Antioxidant enzymes		
Manganese superoxide dismutase	**C08A9.1**	**17.8**
Catalase	Y54G11A.6	6.4

Table 5.4: Heat shock proteins, UGTs and antioxidant enzymes regulated by DAF-16. No further annotation is available for the UGT family members. Potential direct targets of DAF-16 are shown in **bold.**

contain binding sites for both DAF-16 and HSF-1.

Hsu *et al.* (2003) have identified several heat-inducible genes, known as shsps, or small heat shock proteins, whose expression was affected by mutations in the ILS pathway. The authors noted that both DAF-16 and HSF-1 are required for expression of these genes, which all contain both DAF-16 and HSF-1 binding sequences. RNAi of these genes reduced lifespan. These results suggest that DAF-16 and HSF-1 regulate lifespan, at least in part, by increasing expression of shsps.

In the present study, the only group of genes in which both the HSF-1 and DAF-16 binding sites have been detected as over-represented is Mount 15, emphasising the important role of these unknown genes. (However, in Mount 15 only over-representation

of the vertebrate HSF-1 site, but not of the *C. elegans* heat shock element, was detected). Unexpectedly, the *C. elegans* HSE itself appears to be associated with indirect DAF-16 target gene classes. However, the results of the current study depend upon the particular functional groups of genes analysed. To more fully investigate the presence of genes controlled by both factors, analysis of individual genes, or of groups containing only direct DAF-16 target genes, will be necessary.

The *C. elegans* HSF-1 gene lacks a DAF-16 binding element and HSF-1 does not appear to be transcriptionally regulated in DAF-2 mutants. However, activation of HSF-1 is thought to be post-translational (DiDomenico *et al.*, 1982), so it is possible that DAF-16 may activate HSF-1 indirectly. This would lead to concurrent activation of the two factors, permitting them to act together to increase lifespan.

UGTs Consistent with the overall ageing-promoting role of UGTs, 16 UGT family members were down-regulated and 9 were up-regulated in DAF-2 mutants. This work therefore indicates that specific UGTs are likely to play a role in extension of lifespan, some as direct DAF-16 targets and others as indirect targets. Once the UGT family has been more fully characterised it will be interesting to compare the substrates and reactions catalysed by ageing- and longevity-associated UGTs, in an attempt to identify the mechanisms by which some UGTs may extend lifespan.

Antioxidant enzymes While antioxidant enzymes did not rank in the top 20 most over-represented EASE classes, GO0006801 (superoxide metabolism) was significantly over-represented (P=0.05). (The lack of over-representation of other ROS metabolism families may be due to bias towards large gene groups (D. Gems, unpublished observation)). In addition, both catalase and manganese superoxide dismutase (SOD) were strongly up-regulated, suggesting a role for these enzymes in longevity, by protecting cells from oxidative damage. The up-regulation of SOD in DAF-2 mutants has previously been noted (Honda & Honda, 1999).

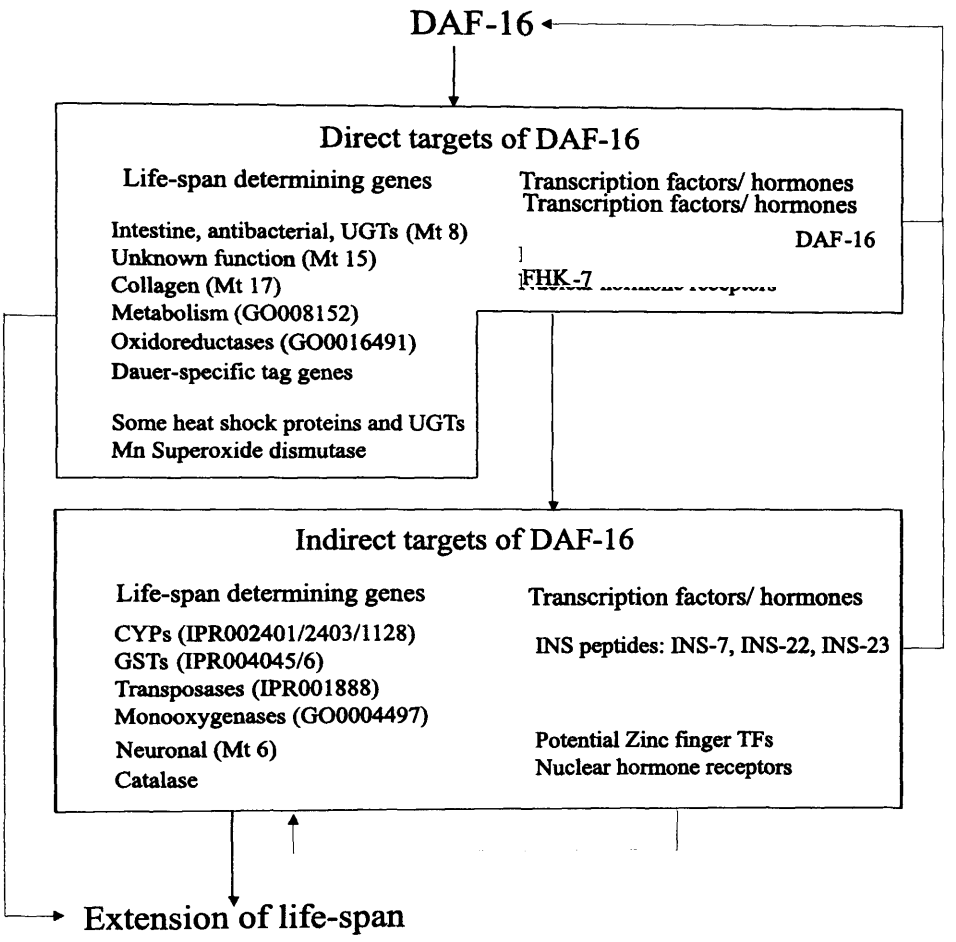


Figure 5.5: Hypothesised mechanism by which DAF-16 regulates lifespan, showing direct and indirect DAF-16 targets, as identified in this study. 'Mt' refers to topomountain groups (Section 5.2.4). See Sections 5.3.4 and 5.3.3 for a discussion of FHK-7, nuclear hormone receptors, insulin-like (INS) peptides and potential zinc fingers.

5.3.1.4 Longevity-associated transcription factors

Clover was used to identify known TFBSs that are over-represented amongst the 20 longevity-associated gene classes. Table 5.5 shows the TFBS motifs that were significantly over-represented in each gene set analysed. These are the known motifs that are most likely to be responsible for the expression patterns observed. It must be remembered that only TFs with their binding site represented in Transfac were identified by this method, and that these TFBSs may represent only a small subset of the actual TFs involved in the control of these sets of genes.

TFBS	Raw score	P	Location
Genes two-fold up-regulated on dauer exit			
AML	36	0.002	5prime
AR	52	0.005	5prime
CREB	6	0	5prime
DAE	3	0.002	5prime
DBE	7	0	5prime
FAC1	117	0	5prime
fDBE	1	0.003	5prime
FOX	27	0	5prime
GR	25	0	5prime
IRF	113	0	5prime
LEF1	25	0.002	5prime
MAF	12	0	5prime
Nrf	2	0.007	5prime
Pbx	2	0.008	5prime
SMAD	10	0.009	5prime
SRY	39	0	5prime
TATA	7	0	5prime
Zeste	8	0.007	5prime

Table 5.5: *Continued on next page*

TFBS	Raw score	P	Location
Genes four-fold up-regulated on dauer exit			
AML	7	0.002	5prime
AP-1	2	0.008	5prime
DAE	4	0.002	5prime
DBE	5	0	5prime
FAC1	65	0	5prime
FOX	24	0	5prime
GR	12	0.001	5prime
LEF	16	0.002	5prime
SMAD	10	0.001	5prime
SRY	43	0	5prime
STAT6	33	0.006	intron1
TATA	9	0	5prime
Genes eight-fold up-regulated on dauer exit			
AML	5	0	5prime
AP-1	4	0	5prime
c-Myc:Max	7	0.003	5prime
DAE	4	0	5prime
DBE	2	0.005	5prime
DBE	-1	0.006	intron1
FAC1	31	0.003	5prime
fDBE	1	0.007	5prime
FOX	13	0	5prime
GR	8	0.004	5prime
PAX	6	0.009	intron1
Pbx	3	0.008	5prime
SRY	13	0	5prime
TATA	6	0	5prime

Table 5.5: *Continued on next page*

TFBS	Raw score	P	Location
Class 1 genes: Longevity-associated genes			
AhR	1	0.005	intron1
Croc	2	0.001	5prime
DAE	21	0	5prime
DBE	15	0	5prime
FAC1	53	0.009	5prime
fDBE	8	0	5prime
FOX	25	0	5prime
GR	10	0.006	5prime
Grainyhead	15	0.003	5prime
HFH	3	0.006	5prime
IRF	77	0.008	5prime
mtTFA	1	0	5prime
PAX	2	0.002	intron1
SGF	8	0.009	5prime
SRY	27	0	5prime
T3R	6	0	5prime
TII	3	0.001	intron1
Mount 6: Neuronal			
Elk-1	20	0.007	5prime
FOX	21	0.004	5prime
ISRE	16	0.008	5prime
LEF	20	0.003	5prime
PAX	7	0	intron1
PPARG	0.1	0.003	5prime
STAT	50	0.001	5prime

Table 5.5: *Continued on next page*

TFBS	Raw score	P	Location
Mount 8: Intestine, antibacterial, UGTs			
ACAAT	6	0.008	5prime
AP-1	4	0	5prime
ARP-1	7	0.009	5prime
Croc	1	0.002	5prime
DAE	23	0	5prime
DAE	-2	0.009	intron1
DBE	21	0	5prime
fDBE	10	0	5prime
FOX	34	0	5prime
GR	12	0.002	5prime
MAF	5	0.008	5prime
MtTFA	1	0	5prime
SRY	30	0	5prime
SRY	7	0.001	intron1
STAT	39	0.003	intron1
TATA	0.4	0.009	5prime
Zta	0.6	0.009	5prime
Mount15: No associated function			
Arnt	4	0.006	5prime
fDBE	3	0.001	5prime
DAE	3	0	5prime
DBE	6	0	5prime
FAC1	46	0.007	5prime
FOX	15	0.001	5prime
GATA	26	0.001	5prime
Hand1:E47	8	0.001	5prime
HFH	2	0.001	5prime

Table 5.5: *Continued on next page*

TFBS	Raw score	P	Location
HSF	14	0.002	5prime
LEF1	15	0.004	5prime
MAF	4	0.008	5prime
MEIS	4	0.002	5prime
MYB	6	0.004	5prime
PAX	4	0.002	intron1
SREBP	3	0.008	5prime
SRY	16	0.008	5prime
USF	17	0.001	5prime
Mount 17: Collagen			
DAE	6	0	5prime
DBE	10	0	5prime
fDBE	5	0	5prime
FOX	10	0.006	5prime
MEF	2	0.003	intron1
NF-Y	2	0.003	intron1
SMAD	1	0.005	intron1
SRY	9	0.008	5prime
SAGE dauer-specific tags			
Cdx	1	0.009	5prime
DBE	2	0.007	5prime
E2A	9	0.006	5prime
FOX	11	0.009	5prime
MyoD	11	0.005	5prime
Myogenin	14	0.005	5prime
SREBP	1	0.001	intron1
SRY	8	0.006	5prime

Table 5.5: *Continued on next page*

TFBS	Raw score	P	Location
IPR004045: Glutathione-S-transferase			
FXR/RXR	0.3	0.005	5prime
HSE	13	0	5prime
Ik	4	0.007	5prime
NF-Y	3	0.004	intron1
SMAD	4	0.003	5prime
IPR004046: Glutathione-S-transferase			
HSE	11	0.003	5prime
NF-Y	3	0.003	intron1
SMAD	4	0.009	5prime
GO0008152: Metabolism			
DBE	2	0.005	5prime
FOX	12	0	5prime
GCNF	1	0.002	intron1
HFH	1	0.005	5prime
Ik	2	0.003	5prime
MEF	0	0.003	intron1
SREBP	0	0	intron1
SRY	11	0.003	5prime
IPR002401: E-class CYPs, group I			
AML	6	0	5prime
c-Myc:Max	2	0.004	intron1
FAC1	14	0.007	5prime

Table 5.5: *Continued on next page*

TFBS	Raw score	P	Location
TGIF	4	0.007	5prime
IPR002403: E-class CYP, group IV			
AML	6	0	5prime
FOX	1	0.009	intron1
Ik	2	0.006	5prime
Nkx2.2	0	0.009	intron1
PAX3	1	0.006	5prime
TGIF	3	0.007	5prime
Zen	1	0.004	5prime
IPR001128: Cytochrome P450			
AML	5	0	5prime
c-Myc:Max	2	0.004	intron1
FAC1	14	0.007	5prime
TFIF	4	0.007	5prime
USF	2	0.007	intron1
Cytochrome P450s			
AML	5	0.001	5prime
c-Myc:Max	2	0.009	intron1
FAC1	14	0.005	5prime
Ik	2	0.006	5prime
MAF	1	0.008	intron1
TGIF	5	0.003	5prime
GO0004497: Monooxygenase activity			

Table 5.5: *Continued on next page*

TFBS	Raw score	P	Location
AML	5	0	5prime
c-Myc:Max	2	0.004	intron1
FAC1	14	0.007	5prime
TGIF	4	0.007	5prime
USF	2	0.007	intron1
IPR001888: Transposases type 1			
AR	2	0.001	5prime
HSAS	4	0.003	5prime
HSE	8	0	5prime
HNF	3	0.001	5prime
MEIS	5	0.001	5prime
NF-AT	7	0.006	5prime
NF-kappaB	6	0.004	5prime
PAX3	4	0.006	5prime
T3R	4	0	5prime
Zic	8	0	5prime
Mariner transposases			
HSE	8	0	5prime
GATA	7	0.006	5prime
GO0016491: Oxidoreductases			
AML	4	0.002	5prime
c-Myc:Max	5	0.009	5prime
fDBE	1	0.008	5prime
DBE	4	0	5prime
FOX	8	0	5prime

Table 5.5: *Continued on next page*

TFBS	Raw score	P	Location
GCNF	2	0.002	intron1
GR	7	0.005	5prime
HFH	3	0.002	5prime
Ik	2	0.008	5prime
MEF	1	0	intron1
NF-Y	2	0.006	intron1
SRY	8	0.002	5prime

Table 5.5: Potential longevity-associated TFBSs, identified by Clover. These motifs were significantly over-represented (P<0.01) in longevity-associated genes. Multiple motifs belonging to the same family of TF were removed to give what is thought to be a non-homologous list of longevity-associated factors. The DAF-16, DBE, DAE, HSE and HSAS motifs (see Figure 5.4) are all derived from *C. elegans* sequences. All other motifs are of vertebrate, mainly mammalian, origin, taken from Transfac.

Transfac contains very few *C. elegans* TFBSs and many of the sites identified here have been found using matrices derived from humans or mice. This is possible since the binding specificities of TFs are generally well conserved between species. Frith *et al.* (2004) have shown that Clover is effective at identifying motifs, even when the motif library used contains only a homologue of the actual binding factor. Hence, the finding that a large number of mammalian sites are over- or under-represented in the *C. elegans* sequences indicates that similar motifs are present and functional.

The main limitation of the use of TFBS matrices from higher organisms to study *C. elegans* is the difficulty in precisely identifying the binding factor. It should be emphasised that Clover identifies over-represented sites, but does not directly identify the binding factor. The latter can only be inferred indirectly from annotations within the motif library. In many cases, such as the glucocorticoid receptor (Section 5.3.1.5), orthology is not detectable and it is not clear which *C. elegans* factor will bind to this site or whether it is valuable to inherit functional information from the mammalian factor to the nematode.

Throughout this chapter, where the role of the TF is known in higher organisms, this is used to hypothesise a possible mechanism by which the factor may control lifespan

in *C. elegans*. It is hoped that in the majority of cases this is justified, even where *C. elegans* lacks the specific target organ or proteins of the TF concerned. For example, while *C. elegans* lacks a heart, in the nematode TFs involved in heart development, such as Nkx2.5, are thought to control development of another peristaltic organ, the pharynx. However, the problems of orthology must be considered when interpreting this data. More work is required to more clearly define orthologous relationships and to test the functional hypotheses proposed here.

Ageing- and longevity-associated TFBSs were identified and compared, as shown in Figure 5.6. Longevity-associated TFBSs are enriched in genes that are up-regulated in DAF-2 mutants. Conversely, ageing-associated TFBSs are under-represented in DAF-16 up-regulated genes, or are over-represented in the genes down-regulated in DAF-2 mutants. TFBSs over-represented in the ageing-associated genes are described in Section 5.3.2. Binding sites associated with the longevity-promoting genes are the DBE, other FOX elements, FAC-1, AML, SRY, the glucocorticoid receptor, c-Myc:Max, IK-2 and TGIF. The interpretation of these data may contribute to our understanding of the role of DAF-16 in control of lifespan, and is described in this section.

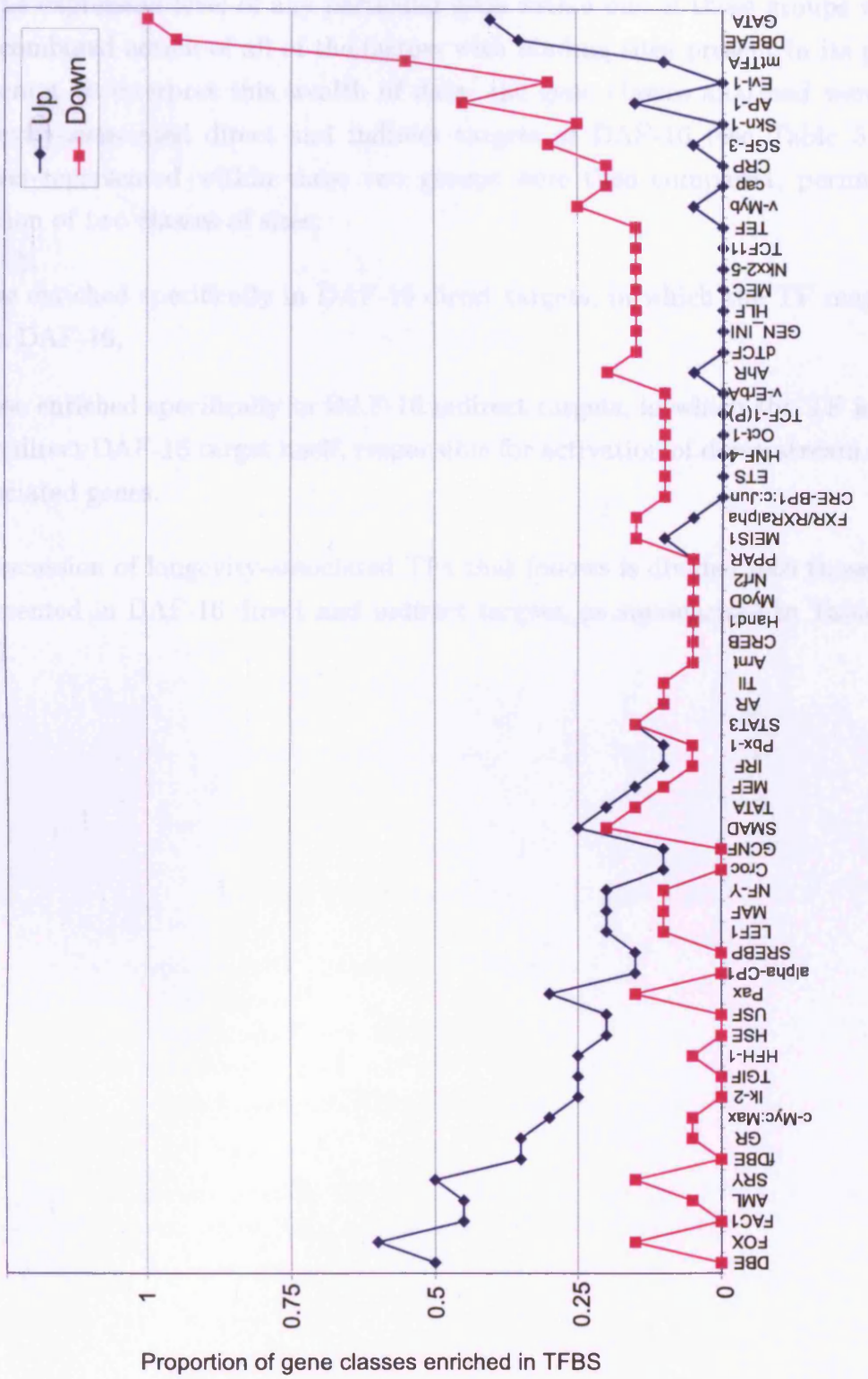


Figure 5.6: Chart showing that certain TFBSs are associated with longevity and with ageing. These sites are over-represented in up- and down-regulated genes in DAF-2 mutants respectively. TFs were excluded from this figure if they were over-represented in only a single ageing- or longevity associated gene class. The y-axis represents the proportion of the 20 ageing- or longevity-associated gene classes in which each factor is significantly over-represented.

As can be seen in Table 5.5, each gene class generally contains many over-represented motifs. The expression level of any particular gene within one of these groups will result from the combined action of all of the factors with binding sites present in its promoter. In an attempt to interpret this wealth of data, the gene classes analysed were divided into longevity-associated direct and indirect targets of DAF-16 (see Table 5.6). The TFBSs over-represented within these two groups were then compared, permitting the identification of two classes of sites:

- those enriched specifically in DAF-16 direct targets, in which the TF may interact with DAF-16,
- Those enriched specifically in DAF-16 indirect targets, in which the TF is a candidate direct DAF-16 target itself, responsible for activation of down-stream longevity associated genes.

The discussion of longevity-associated TFs that follows is divided into those that are over-represented in DAF-16 direct and indirect targets, as summarised in Table 5.6 and Figure 5.7.

Factor	General function
	Direct DAF-16 targets
SRY	Development of male gonads and sexual differentiation
GR	Stress, diverts metabolic resources
	Indirect DAF-16 targets - All
HSE	Heatshock
TGIF	Repressor of RXR- and TGF-β-dependent transcription
cMyc-Max	Proliferation
USF	Inhibits proliferation, stress response
	Indirect DAF-16 targets - With HSE
HSE	Heat shock response
	Indirect DAF-16 targets - No HSE
AML-1	Haematopesis (leukaemia), proliferation
cMyc:Max	Proliferation
FAC1	Brain development and injury response
TGIF	Repressor of RXR- and TGF-β-dependent transcription
USF	Inhibits proliferation, stress response

Table 5.6: Transcription factor binding sites over-represented in longevity-associated genes. See text for references and further discussion.

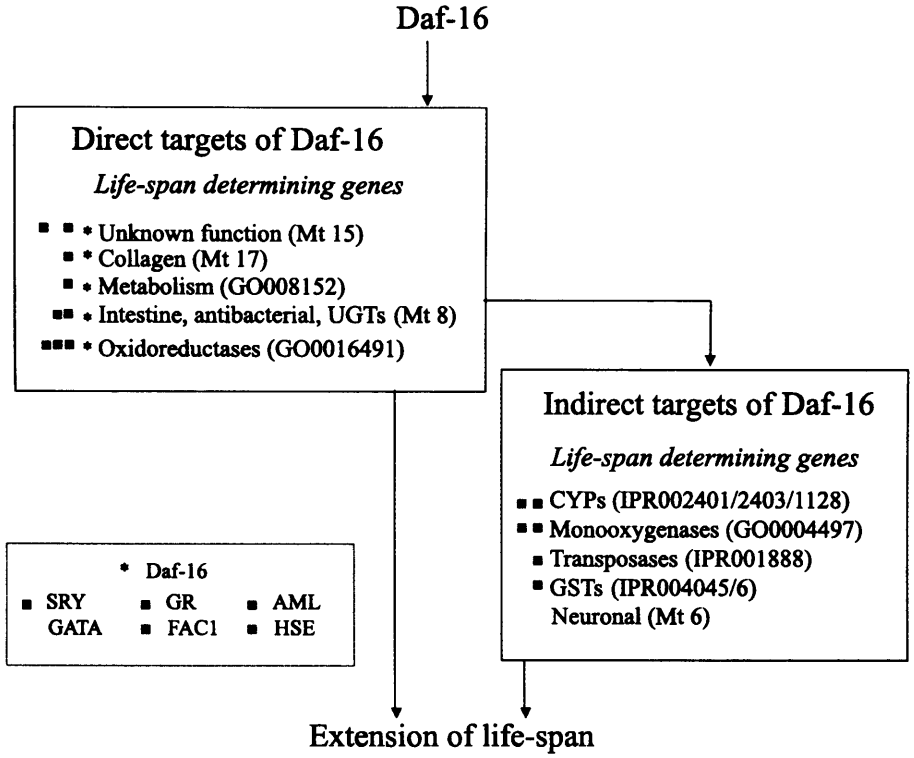


Figure 5.7: Hypothesised mechanism by which DAF-16 regulates lifespan, showing direct and indirect DAF-16 targets and some important over-represented TFBSs, as identified in this study.

5.3.1.5 Longevity-associated transcription factor binding sites (over-represented in direct targets)

As shown in Figure 5.8, certain TFBSs are associated with direct transcriptional targets of DAF-16, while others tend to be associated with indirect targets. The association of another TFBS with DAF-16 suggests that either:

- Levels of the associating TF are increased in long-lived animals and the two TFs physically interact to control the same or similar genes. In this case the TFBS is enriched in the longevity-associated genes for specific functional reasons.
- The two TFs have a similar function and therefore independently control the same genes under different conditions. In this case, the TFBS is simply enriched by chance, and expression of the TF is unlikely to be increased in DAF-16 mutants.

Some attempt will be made to distinguish between these two possibilities in cases where the *C. elegans* homologue is known and has associated expression data. In either case, it will be likely that some of the known targets of the associated TF will be longevity-determining. It is therefore informative to study the function and targets of these TFs, in order perhaps to learn more about the mechanisms by which lifespan is controlled.

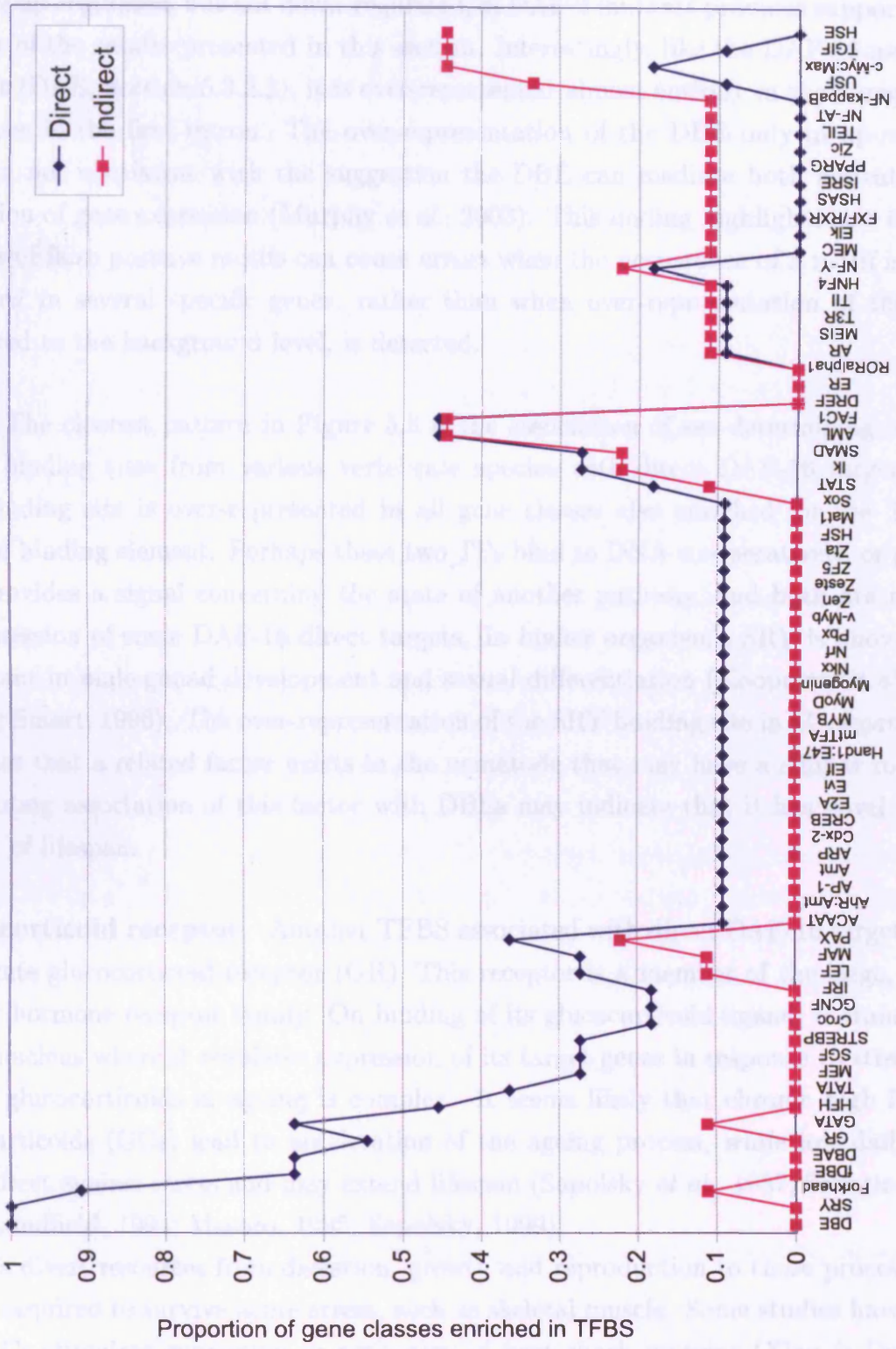


Figure 5.8: Chart showing that certain TFBS are associated with direct transcriptional targets of DAF-16, while others tend to be associated with indirect targets. The y-axis represents the proportion of direct or indirect target gene classes in which a particular TFBS is over-represented.

DAF-16 binding element The fact that the DBE is strongly over-represented in genes that are up-regulated, but not down-regulated, in DAF-2 mutants provides support for the validity of the results presented in this section. Interestingly, like the DAF-16-associated element (DAE, Section 5.3.2.2), it is over-represented almost entirely in the 5' region and only once in the first intron. The over-representation of the DBE only in up-regulated genes in not consistent with the suggestion the DBE can mediate both activation and repression of gene expression (Murphy *et al.*, 2003). This finding highlights how the large number of false positive motifs can cause errors when the occurrence of a motif is simply identified in several specific genes, rather than when over-representation of the motif, compared to the background level, is detected.

SRY The clearest pattern in Figure 5.8 is the association of sex-determining region Y (SRY) binding sites from various vertebrate species with direct DAF-16 targets. The SRY binding site is over-represented in all gene classes also enriched for the Transfac DAF-16 binding element. Perhaps these two TFs bind to DNA co-operatively, or perhaps SRY provides a signal concerning the state of another pathway, and both are required for expression of some DAF-16 direct targets. In higher organisms, SRY is known to be important in male gonad development and sexual differentiation (Koopman *et al.*, 1990; Shah & Smart, 1996). The over-representation of the SRY binding site in *C. elegans* genes indicates that a related factor exists in the nematode that may have a similar role. The very strong association of this factor with DBEs may indicate that it has novel roles in control of lifespan.

Glucocorticoid receptor Another TFBS associated with direct DAF-16 targets is the vertebrate glucocorticoid receptor (GR). This receptor is a member of the large, diverse nuclear hormone receptor family. On binding of its glucocorticoid ligand, it translocates to the nucleus where it regulates expression of its target genes in response to stress. The role of glucocorticoids in ageing is complex. It seems likely that chronic high levels of glucocorticoids (GCs) lead to acceleration of the ageing process, while low daily levels can protect against stress and may extend lifespan (Sapolsky *et al.*, 1987; Sabatino *et al.*, 1991; Landfield, 1994; Masoro, 1995; Sapolsky, 1999).

GCs divert resources from digestion, growth and reproduction to those processes and tissues required to survive acute stress, such as skeletal muscle. Some studies have shown that GCs stimulate expression or activation of heat shock proteins (Xiao & DeFranco, 1997; Vijayan *et al.*, 2003; Nedellec *et al.*, 2002), where as others have shown the opposite result (Wadekar *et al.*, 2001; Boone & Vijayan, 2002; Wadekar *et al.*, 2004). This is perhaps due to differences in the roles of particular heat shock proteins, or perhaps a

result of experimental inconsistencies. Among the targets of the GR are known to be the cytochrome P450, CYP2B2, (Jaiswal *et al.*, 1990) and the human growth hormone gene (Moore *et al.*, 1985).

It is conceivable that these and other functions of glucocorticoids promote lifespan extension by protecting against stress induced damage and diverting resources to maintenance and repair. Consistent with this is the finding that stress can potentiate the action of glucocorticoids (Sanchez *et al.*, 1994; Hu *et al.*, 1996; Li *et al.*, 2000; Jones *et al.*, 2004). Also, interestingly, basal levels of GCs are considerably elevated in aged rats (Sapolsky, 1992).

C. elegans lacks a clear orthologue of the mammalian GR (Maglich *et al.*, 2001). However, the over-representation of GR-like binding sites in direct DAF-16 targets indicates that a related nuclear hormone receptor exists that is likely to play a role in ageing. Given the similarity of the binding site of this unknown factor to the mammalian GR binding site, and the potential role of the GR in mammalian ageing, it is tempting to postulate that the two factors may have a similar role. The over-representation of the GR-like binding sites in direct DAF-16 targets indicates that there may be an overlap in the target genes and mechanisms of action by which these two TFs extend lifespan.

An interesting consideration is why SRY and GR binding sites are found almost exclusively in direct, but not indirect, DAF-16 targets. One hypothesis could be that the SRY-like and GR-like factors physically interact with DAF-16 on DNA binding. Alternatively, perhaps the classes of genes regulated directly by DAF-16 are fundamentally different in function to the indirect targets. For example, perhaps the indirect targets of DAF-16 are involved in aspects of longevity that it would be inappropriate to regulate by SRY and GR, such as dauer specific functions. Considerable further work will be needed to fully understand the interactions between these factors in the control of lifespan. It will be particularly important to investigate the impact that the presence of a dauer phase has on the mechanisms used to regulate longevity in different species.

5.3.1.6 Longevity-associated transcription factor binding sites (over-represented in indirect targets)

Binding sites associated with indirect up-regulated DAF-16 targets can provide insight in two main ways:

- For cases in which there is considerable knowledge concerning the transcriptional targets of the TF, further clues concerning the mechanism by which lifespan is extended can be obtained. Knowledge of the interaction of such factors with coactivators or corepressors can provide information concerning the regulation of longevity

and the integration of pathways controlling it.

- Where the role of the TF concerned is relatively poorly understood, the finding that it is associated with genes up-regulated by DAF-16 can suggest that its function may include longevity-promoting actions.

A number of TFs belonging to each of the above classes are discussed below.

There is no single TFBS over-represented in all indirect target classes of DAF-16. However, the heat shock element appears to be associated mainly or entirely with the indirect targets. The HSE is over-represented in 4 of the 10 indirect target gene classes. 5 of the 6 remaining gene groups are all related to CYPs and all show very similar patterns of TF over-representation. As is shown in Figure 5.9, they are all strongly associated with UFS-1, FAC-1, AML-1, TGIF and c-Myc:Max binding sites. The final indirect target gene class is Mount 6, neuronal genes, which shows very different patterns of TFBS over-representation. Hence, longevity-associated genes can be divided into the following groups:

- Direct DAF-16 targets
- Indirect DAF-16 targets
  - HSE-containing indirect targets
  - Cytochrome P450s
  - Neuronal genes (Mount 8)

Heatshock element The *C. elegans* heat shock element was identified by GuhaThakurta *et al.* (2002) in the promoters of *C. elegans* genes up-regulated in response to heat shock. This motif appears to be involved in the control of expression of indirect DAF-16 targets. Heat shock elements are discussed in more detail in Section 5.3.1.3.

TGIF TGIF is a transcriptional repressor which is thought to act both by remodelling of chromatin structure and competition for DNA binding sites with activators (Wotton *et al.*, 1999b). For example, TGIF is known to inhibit transcription from retinoid-X-receptor (Bertolino *et al.*, 1995) and TGF-β dependent promoters (Wotton *et al.*, 1999a).

The mechanism by which these or other actions of TGIF may contribute to extension of lifespan is not currently understood. That TGIF is thought to act as a repressor but its site is enriched in up-regulated genes may argue against a direct role in longevity.

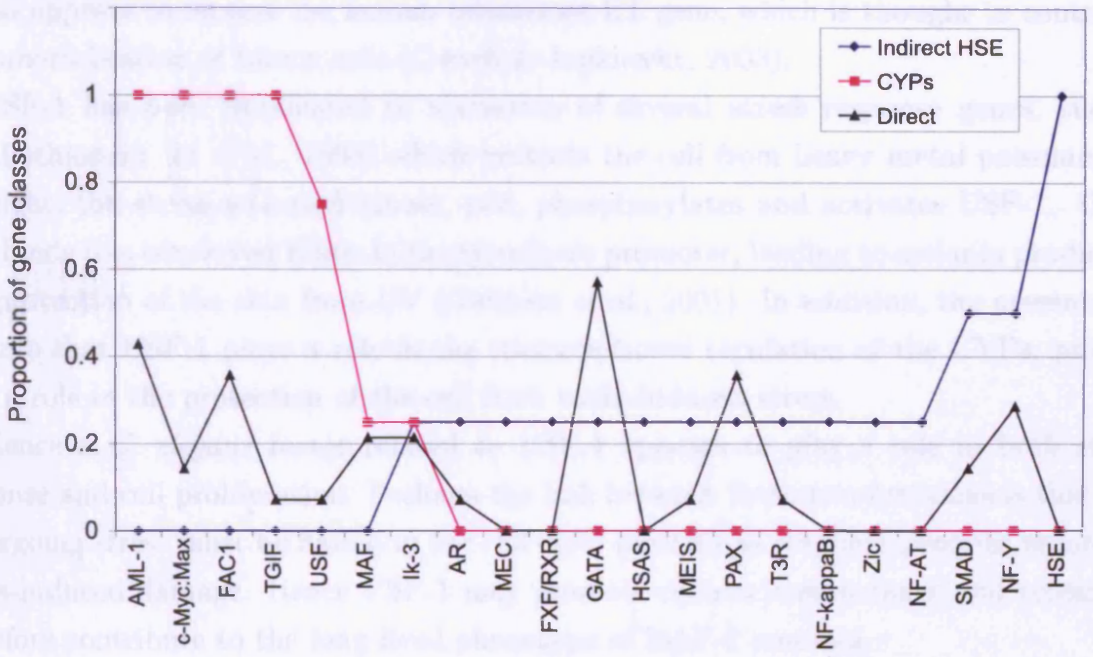


Figure 5.9: Chart showing that certain TFBS are associated with indirect transcriptional targets of DAF-16 enriched for the HSE, while others tend to be associated with CYPs (indirect targets lacking the HSE). The profile for direct DAF-16 targets is also given, for comparison. The y-axis represents the proportion of gene classes in which a particular TFBS is over-represented.

However, the finding that CYP promoters are highly enriched in TGIF-like binding sites requires further investigation.

c-Myc:Max c-Myc:Max is a heterodimeric protooncogene (Amati *et al.*, 1993), involved in cellular transformation, proliferation and apoptosis (Evan & Littlewood, 1993; Henriksson & Luscher, 1996). It has been shown to activate p48, a stress response factor that reduces the sensitivity of the tumour cells to anti-cancer drugs (Weihua *et al.*, 1997).

Mammalian c-Myc:Max binding sites, known as E boxes, are over-represented in the non-HSE containing indirect DAF-16 targets. While a *C. elegans* homologue for c-Myc has not been identified, MXL-1 is a good candidate for the homologue of Max (Yuan *et al.*, 1998). Clearly, empirical studies are required to confirm the identity of the *C. elegans* TF that is binding to these sites. Alternatively, c-Myc:Max binding sites may appear to be over-represented due to their high similarity to the over-represented USF-1 binding site.

USF-1 Many non-HSE-containing gene classes are enriched for the upstream stimulatory factor 1 (USF-1) binding site, also an E box. USF-1 binds to the E box, limiting

the stimulatory effect on the factor TFE-3 on cell proliferation (Kiermaier *et al.*, 1999). It also appears to repress the human telomerase RT gene, which is thought to contribute to immortalisation of tumor cells (Goueli & Janknecht, 2003).

USF-1 has been implicated in activation of several stress response genes, such as metallothionein (Li *et al.*, 1998) which protects the cell from heavy metal poisoning. In UV light, the stress-activated kinase, p38, phosphorylates and activates USF-1. USF-1 then binds to a conserved E box in the tyrosinase promoter, leading to melanin production and protection of the skin from UV (Galibert *et al.*, 2001). In addition, the present data indicate that USF-1 plays a role in the transcriptional regulation of the CYPs, proteins with a role in the protection of the cell from toxin-induced stress.

Hence a *C. elegans* factor related to USF-1 appears to play a role in both stress-response and cell proliferation. Perhaps the link between these two functions is that cells undergoing stress must be halted in the cell cycle until stress response proteins repair any stress-induced damage. Hence USF-1 may promote cellular maintenance and repair and therefore contribute to the long-lived phenotype of DAF-2 mutants.

Whelan *et al.* (1990) have shown that USF is able to bind to the insulin control element, upstream of the rat insulin II gene. While they found that USF binding did not stimulate insulin transcription, they could not rule out that USF may negatively regulate transcription of insulin. If USF does inhibit insulin II transcription, this may represent another longevity-promoting action, since reduced insulin like signalling leads to activation of DAF-16.

AML-1 AML-1 (acute myeloid myelogenous leukaemia-1) TF is involved in the control of haematopoietic cell proliferation (Ichikawa *et al.*, 2004) in higher organisms. Over-representation of AML-1 binding sites in the CYP gene promoters suggests that a related *C. elegans* factor exists that may co-stimulate haematopoesis and CYP expression. This may be valuable in times of infection, for example, in order to promote both an immune response and detoxification of bacterial toxins.

FAC-1 FAC-1 is a developmentally regulated TF, of which levels are high in the developing brain, but significantly lower in adults (Bowser *et al.*, 1995). Re-expression of FAC-1 in adults occurs during brain injury and some neurodegenerative diseases (Styren *et al.*, 1997), suggesting that FAC-1 mediates a repair response to neuronal damage. In these situations, FAC-1 may additionally stimulate the CYPs, as found in this work, in order to protect the cell from endogenous toxins, released from damaged cells.

Summary In most cases a possible explanation for the over-representation of AML-1, FAC-1, c-Myc:Max, TGIF and USF-1 binding sites in longevity-associated genes can be tentatively proposed, based on what is known of their function in higher organisms. It will be important to identify the *C. elegans* factors that bind to these sites and determine their cellular functions. This work has lead to the generation of many hypotheses that now require substantial further testing to more clearly understand the role of these TFBSs in longevity. The study has also illustrated the tremendous level of complexity with which numerous TFs interact to control and integrate physiological processes such as stress responses, cell proliferation and longevity.

5.3.2 Ageing-associated genes and their regulation

5.3.2.1 The role of ageing-associated gene groups in the control of lifespan by DAF-16

There are a number of functional themes associated with the ageing-promoting gene classes. These themes are discussed in the following sections:

Transporters Reduced expression of transporter genes (IPR005828, GO006810, transporters) in DAF-2 mutants may contribute to their long-lived phenotype by reducing nutrient uptake in the intestine and inducing dietary restriction (McElwee *et al.*, 2004).

Amino acid and lipid metabolism A large number of ageing-associated gene groups have roles in metabolism of lipids and amino acids (Mounts 19, 24, 27 and 21 and 'lipid metabolism genes'). However, the mechanism by which these genes may affect lifespan remains poorly understood. The function of this down-regulation may be to reduce protein and lipid synthesis, as resources are switched from growth and reproduction to maintenance. In support of this idea, lipid metabolism is known to be reduced, and lipid storage increased, in dauers (O'Riordan & Burnell, 1990).

UDP-glucuronosyltransferases and Mount 8 As described previously, McElwee *et al.* (2004) observed that UGTs were up-regulated in dauers, along with other possible xenobiotic protective groups, and hypothesised that these genes may contribute to the longevity of dauer larvae. However, while the other xenobiotic detoxification gene classes are up-regulated in DAF-2 mutants, the UGTs are down-regulated (IPR000213 and the UGTs).

McElwee *et al.* (2004) suggested that the down-regulation of UGTs in DAF-2 mutants but not dauers is related to differences in the targets of detoxification in adults and dauer

larvae. As described in Section 5.3.1.3, a small number of individual UGT genes were up-regulated in the McElwee *et al.* (2004) study and UGT up-regulation was observed in a previous study (Murphy *et al.*, 2003). Taken together, these data suggest that some UGTs are up-regulated in DAF-2 mutants and may contribute to longevity via a protection from chemical stress. However, other UGTs may promote ageing, perhaps through the generation of toxic metabolites.

In addition some components of Mount 8 are up-regulated and some are down-regulated in DAF-2 mutants. Given the strong down-regulation of the other UGT groups it seemed possible that this function may also account for the partial down-regulation of Mount 8. However, while no up-regulated UGT family members were present, only 3 out of 16 down-regulated UGT family members were found in Mount 8. This indicates that other functional groups of genes are contribute to the down-regulation of Mount 8. One possibility is that some down-regulated Mount 8 genes may be intestinal genes involved in digestion and nutrient uptake. Down-regulation of these genes would be predicted to extend lifespan by inducing dietary restriction.

C-lectins C-type lectins are a large, variable family of Ca^{2+}-dependent carbohydrate binding proteins with no known catalytic activity. While the functions of specific lectins are poorly understood, some studies have suggested a role in immunity (Rosen, 1993; Lasky, 1995) and cell adhesion (Lasky, 1994). Why immune proteins would be down-regulated in long-lived organisms is unclear, since increased immunity would appear to protect the cell from infection and so prolong life.

Other ageing-associated gene groups In addition, it will be important to characterise the genes within the 2- and 4-fold down-regulated classes, and the class 2 genes, that do not belong to one of the above categories (see Table 5.2). This may allow further patterns of DAF-2 mutant gene expression to be defined. Similarly, further insights may come when the 'Domain of Unknown Function' (DUF) gene families (DUF 141, 227 and 274) are functionally characterised, particularly since RNAi of DUF141 genes has been shown to extend lifespan (Murphy *et al.*, 2003).

5.3.2.2 Ageing-associated transcription factor binding sites (over-represented in down-regulated genes)

Ageing-associated transcription factor binding sites (over-represented in down-regulated genes) are listed in Table 5.7. The actions of these TFs are likely to be linked to pathways that accelerate ageing, including growth, reproduction and inhibition of damage repair mechanisms.

Factor	General function
DAE	Unknown. Previously associated with ageing-promoting genes
GATA	Development
mtTFA	Mitochondrial gene transcription
Evi-1	Development, cell proliferation and differentiation
AP-1	Stress response
SKN-1	Development of and stress responses in the digestive tract

Table 5.7: Transcription factor binding sites over-represented in ageing-associated genes. See text for references and further discussion.

The DAF-16 associated element The DAF-16 associated element (DAE) was first identified by Murphy *et al.* (2003). Using two different computational approaches the group showed that the sequence CTTATCA is over-represented in both up- and down-regulated genes in DAF-2 mutants. This sequence is here used to generate an 'artificial' TFBS matrix for use with Clover. However, the DAE has more recently been proposed to function exclusively as an ageing-associated regulatory motif (McElwee *et al.*, 2004). The current study, using more advanced statistical techniques, confirms the important role of the DAE in both ageing-and longevity-associated genes, as first described by Murphy *et al.* (2003). Interestingly, like the DBE (Section 5.3.1.5), it is over-represented almost entirely in the 5' region and only once in the first intron.

The DAE is over-represented in 19 out of 20 ageing-associated gene classes and 7 out of 10 up-regulated, direct targets of DAF-16, but in none of the 10 indirect targets. This suggests that the presence of the DAE recruits an as yet unknown factor, which inhibits gene transcription. However, in direct DAF-16 targets, the DBE present recruits DAF-16, which may act to prevent the action of the inhibitory factor bound to a DAE. An interesting way to begin to investigate this hypothesis would be to determine the average distance between the DBE and DAE sites. If the two motifs are generally found close on the DNA, this may suggest that the two factors may heterodimerise, or otherwise physically interact whilst bound, leading to activation of transcription. Conversely, homodimerisation or binding of a monomeric form of the DAE-binding factor may lead to inhibition of transcription.

GATA GATA sites are under-represented in the majority of direct DAF-16 target gene classes (IPR001128, Mount 8, Mount 17, two- and four-fold up and class 1). In contrast, the site is over-represented in heat shock genes, mariner transposases and Mount 15 groups. This result illustrates the complexity and variety of different functions performed

by just a single TF family. GATA TFs are involved in the development of many tissues, particularly the mesendoderm and blood (Patient & McGhee, 2002; LaVoie, 2003).

It is possible that the enrichment of GATA in down-regulated genes is, at least in part, due to a high similarity between the GATA binding consensus and the reverse complement of the DAE. It will be important to more fully characterise the binding sites of the *C. elegans* GATA factors and the sequence of the DAE, in order to determine whether this is true.

The ageing-associated GATA binding sites identified is this work are GATA1 and GATA2 (and to a lesser degree GATA3 and GATA6). This pattern is slightly different to that of GATA factors associated with longevity genes, which were GATAs 1, 2 and 4. However, these motifs are derived from vertebrates and there are no clear orthology relationships between the 6 vertebrate GATAs and the 11 *C. elegans* factors (Patient & McGhee, 2002). Hence at present it is very difficult to interpret this information in order to identify differences in function between the ageing- and longevity-associated GATA factors.

Evi-1 Evi-1 (Ectopic viral integration site 1) is a zinc finger TF involved in development, cell proliferation and differentiation (Morishita *et al.*, 1990; Perkins *et al.*, 1991; Garriga *et al.*, 1993; Hirai, 1999; Hirai *et al.*, 2001). Evi-1 represses the growth-inhibiting actions of TGFβ (Kurokawa *et al.*, 1998). This action might be associated with ageing, since resources must be diverted from maintenance and repair to growth. If the other activities of Evi-1 targets are similar, this may explain the over-representation of Evi-1 binding sites in ageing-associated genes. Unfortunately, the expression in DAF-2 mutants of egl-43, the *C. elegans* homologue of Evi-1, was not assessed in the McElwee *et al.* (2004) study. However, like Evi-1, it is known to be involved in development (Garriga *et al.*, 1993).

AP-1 Expression of AP-1 is stimulated by heat shock (Diamond *et al.*, 1999), oxidative stress or by various chemicals that alter the cellular redox balance (Rahman *et al.*, 2002). It then activates expression of genes involved the response to stress (Gius *et al.*, 1999), including proinflammatory agents and antioxidants. Since AP-1 appears to have longevity-promoting function it is surprising that mammalian AP-1 bindings sites are over-represented amongst ageing-promoting genes. This finding requires further investigation. It may suggest that the *C. elegans* factor that binds to this site does not share its function with the mammalian AP-1 factor.

SKN-1 SKN-1 is a *C. elegans* TF involved in the development of the mesendodermal tissues, including the digestive tract, and in stress response mechanisms in the intestine (An & Blackwell, 2003). SKN-1 binding sites have been identified upstream of several oxidative and xenobiotic stress resistance genes, including superoxide dismutases (SOD1-3) and catalase, CTL-1 (An & Blackwell, 2003). Like AP-1, it is surprising that a likely longevity-promoting gene such as SKN-1 is over-represented in ageing-associated genes.

Perhaps SKN-1 has both longevity and ageing-promoting roles. Crucially, SKN-1 mutants showed lifespan reduced by 25-30%, demonstrating that SKN-1 is required for normal longevity (An & Blackwell, 2003). The fact that SKN-1 binding sites are not over-represented amongst DAF-16-regulated, longevity-promoting genes indicates that SKN-1 and DAF-16 control lifespan through independent mechanisms. Consistent with this is the finding that, while SKN-1 binding sites are over-represented in heat shock genes, these genes show little regulation by DAF-16.

A possible ageing-associated function of SKN-1 could be in stimulating pharyngeal pumping or expression of intestinal transporter genes. This would increase the rate of nutrient uptake and would therefore be predicted to decrease lifespan. Mutations have been identified that extend lifespan by disrupting pharyngeal function and act by mechanisms independent of DAF-16 (Klass, 1983; Langosch & Heringa, 1998). The potential of SKN-1 in control of lifespan requires further investigation.

5.3.2.3 Ageing-associated transcription factor binding sites (under-represented in longevity-associated genes)

Further potential ageing-associated TFs, under-represented in DAF-16 up-regulated genes, are given in Table 5.8. Under-representation of a TFBS in the direct targets of another TF suggests that the two sets of target genes involve non-overlapping, or even opposing, functions. Hence if a TFBS is infrequently found in longevity-associated genes, that TF may play a role in shortening of lifespan. While very little is known about the function and transcriptional targets of several of these TFs, the current work suggests that they may be involved in diverting resources back from maintenance and repair to reproduction and growth.

Hypoxia inducing factor-1 HIF-1 (Hypoxia-inducible factor-1) mediates an adaptive response to hypoxia (Semenza, 2000, 2004). It is therefore surprising that the *C. elegans* HIF-1 response element (HRE) would be under-represented in DAF-16 targets. The finding indicates that the HIF-1 mediated hypoxic response does not contribute to the long-lived DAF-2 mutant phenotype. Furthermore, since HIF-1 sites are significantly under-represented rather than simply not over-represented, it seems that some of the genes

Factor	General function
HIF-1	Hypoxic response
Nkx2.5	Heart development in vertebrates, possibly development of the pharynx in *C. elegans*
PAX4	Pancreatic development, possibly regulation of insulin expression
Sry-β	Unknown
AREB6	Unknown
CdxA	Unknown
GATA	Development
HNF4	Stress response
HSF	Heat shock and stress responses.
MINI-19	Control of muscle-specific genes
PPAR	Regulation of lipid metabolism
YY-1	Development

Table 5.8: Transcription factor binding sites under-represented in longevity-associated genes. See text for references and further discussion.

involved in the HIF-1 hypoxic response would actually tend to shorten lifespan. Why this may be is unclear. While McElwee *et al.* (2004) did not assess TFBS under-representation, they found that the HRE was significantly over-represented in some longevity-associated and some ageing-associated gene classes. The role of this element in control of lifespan requires further investigation.

Nkx2.5 The vertebrate TF Nkx2.5 is a member of the homeobox family involved in heart development. While *C. elegans* lack a heart, Nkx2.5 appears to be highly functionally related to the *C. elegans* TF CEH-22, which is involved in development of the pharynx (Haun *et al.*, 1998). Ectopic expression of Nkx2.5 in *C. elegans* directly activates expression of CEH-22 target genes, via the CEH-22 binding site. Nkx2.5 can also completely reverse the pharyngeal pumping defects associated with CEH-22 mutations.

While the only known target of CEH-22 is the pharyngeal muscle specific gene *myo-2*, CEH-22 is also thought to regulate the expression of other pharyngeal genes (Okkema & Fire, 1994). Under-representation of CEH-22/Nkx2.5 targets amongst longevity-associated genes implies that these pharyngeal genes may have an ageing-promoting action. This is likely to occur because down-regulation of CEH-22 targets leads to compromised pharyngeal function and extension of lifespan via dietary restriction. Consistent with this idea, CEH-22 mutations often lead to dauer arrest (Haun *et al.*, 1998) and mutation of other pharyngeal genes has been shown to extend lifespan (Klass, 1983; Langosch

& Heringa, 1998).

PAX4 PAX TFs are a widespread family of factors with an important role in development (reviewed in Chi & Epstein (2002)). Each family member appears to play a unique but overlapping role in a specific tissue or set of tissues. All PAX TF bind to a related 'paired domain' sequence. The current analysis has shown that PAX binding sites are under-represented in DAF-16 targets, and are therefore associated with reduction of lifespan.

The role of PAX6 is particularly interesting with respect to the role of insulin-like signalling in control of lifespan. This factor is described as a key regulator of the pancreatic hormones glucagon and insulin as well as pancreatic development (Sander *et al.*, 1997). PAX6 activates insulin transcription, leading to an increase in insulin-like signalling that would be expected to reduce lifespan. There is also some evidence that PAX4 may perform a similar role and that PAX4 may be required to maintain differentiation of the insulin producing β-cells (Smith *et al.*, 1999; Prado *et al.*, 2004). If the other actions of PAX6 and the other PAX factors also accelerate ageing this may explain their under-representation amongst longevity-associated genes.

Serendipity-β Serendipity-β (SRY-β) is a forkhead transcription factor, the role of which is not clearly understood (Payre & Vincent, 1991; Noselli *et al.*, 1992). The finding that its binding sites are under-represented in longevity-associated genes suggests that some of its transcriptional targets may act to accelerate ageing.

AREB6 Human AREB6 binding sites were under-represented in longevity-associated genes. This factor contains two zinc finger motifs and a homeodomain, all of which are thought to contribute to DNA binding. Interestingly the relative contributions of each domain to DNA binding, and the transcriptional outcome, is determined by the presence of a novel motif GTTTC/G (Ikeda & Kawakami, 1995). Using a transfection assay, Ikeda & Kawakami (1995) showed that AREB6 would stimulate expression of human Hsp70 in the absence of this motif, but repress transcription if it was present. This finding suggests that AREB6 may have a role in control of longevity by modulating the heat shock stress response. It will be important to characterise the *C. elegans* factor that binds to this site and to determine whether the GTTTC/G motif is also functional in the nematode.

CdxA While the specific function of this factor has not been established, CdxA belongs to the *Drosophila* homeobox genes, important in embryonic development (Mayinger & Klingenberg, 1992). Interestingly, the CdxA binding site (Margalit *et al.*, 1993) is very

similar to that of the hamster CDX3 factor in the insulin I gene (German *et al.*, 1992), suggesting that CdxA may regulate insulin-like signalling.

GATA See Section 5.3.2.2 for a discussion of the possible roles of GATA family TFs in ageing.

HNF4 Why the HNF4 binding site is under-represented in longevity-associated genes is unclear since the factor has been shown to stimulate expression of a number of longevity-promoting genes. Such targets include glutathione-S-transferase (Paulson *et al.*, 1990), CYPs (Jover *et al.*, 2001) and inducible nitric oxide synthase (iNOS) (Guo *et al.*, 2002). The latter is particularly interesting, since HNF4 has been shown to stimulate iNOS in response to oxidative stress, leading to antioxidant defence. Further work is needed to explain this unexpected under-representation of HNF4 targets in longevity-associated genes. While it is possible that the *C. elegans* factor that binds to these HNF4-like sites has a different or opposing action, the conservation of the motif from *Xenopus laevis* to humans suggests that this is unlikely.

HSF It is unexpected to find the HSF-1 binding site associated with ageing-promoting genes, since heat shock proteins are thought to slow ageing. However, the motif used is derived from vertebrates and it is perhaps recognising another similar but functionally unrelated TF binding site in *C. elegans*. This area requires further investigation.

MINI-19 The muscle initiator sequences-19 (MINI-19) from Transfac Professional v8.1 is a compilation of 7 motifs identified in the promoters of muscle specific genes such as actin and myosin from unspecified species. The binding factor is unknown.

The finding that these structural muscle proteins are under-represented amongst longevity-promoting genes is consistent with the hypothesis that DAF-16 extends lifespan by diverting resources away from growth to maintenance and repair.

PPAR Mammalian PPAR-like binding sites are under-represented in longevity-associated genes. Orthologues of the PPARs have not been found in *C. elegans* indicating that the under-represented site may bind other, non-PPAR, members of the nuclear hormone receptor class of TFs. However, given the significant evidence for a role of PPARs in control of lifespan, it is interesting to discuss the possibility that these sites bind a *C. elegans* PPAR.

PPARs (peroxisome proliferator activated receptors) bind a variety of lipid ligands and stimulate transcription. Interestingly, a polymorphism in the PPARγ gene has been

shown to be linked to longevity (Barbieri *et al.*, 2004). However, while it seems likely that PPARs may contribute to the control of lifespan, PPARα and γ agonists have been shown to promote both ageing- and longevity-associated processes. For example, PPARs are thought to both increase insulin sensitivity (Moller & Berger, 2003; Ma *et al.*, 2004; Shen *et al.*, 2004; Hegarty *et al.*, 2004) and stimulate the expression of the UCPs (Kelly *et al.*, 1998; Medvedev *et al.*, 2001; Son *et al.*, 2001; Grav *et al.*, 2003).

Increases in insulin sensitivity, due to regulation of lipid storage and of other hormones by PPARs, promote increased insulin-like signalling. As a result, the inhibition of DAF-16 by the ILS pathway is strengthened, leading to a reduction in lifespan. Hence ageing-promoting functions of PPAR targets, such as this, may explain the under-representation of PPAR binding sites amongst longevity-associated DAF-16 targets.

However, PPAR-induced expression of UCPs would be predicted to increase lifespan by reducing the production of free radicals (see Chapter 2). In addition, some evidence has emerged to suggest that PPARs may be involved in cell protection (Deplanque, 2004), another action that may extend lifespan. Finally, PPARα is thought to be a key regulator of the CYP4 family and of UGT1A9, stimulating transcription in response to xenobiotic compounds (Waxman, 1999; Gueraud *et al.*, 1999; Barbier *et al.*, 2003).

Hence PPARS may have both ageing- and longevity-promoting actions. Perhaps PPARs promote either ageing or longevity, dependent upon the presence of particular cofactors, which facilitate binding to different sets of target genes. However, due to the lack of overlap between their targets, if PPARs are able to promote longevity, they are likely to function by mechanisms independent of those used by DAF-16.

Yin yang 1 Another of the under-represented motifs is that of the yin yang 1 transcription factor (YY-1). This factor is essential to embryonic development (Donohoe *et al.*, 1999) and is thought to have both activating and repressing actions (Weill *et al.*, 2003). How YY-1 may accelerate lifespan requires further investigation.

5.3.3 Transcription factors regulated by DAF-16

Indirect targets of DAF-16 that are up-regulated in long-lived animals include transposases, oxidoreductases, cytochrome P450s and GSTs. The expression of these genes appears to be affected by a wide range of TFs. However, using TFBlast, the current work has shown that when DAF-16 activity is increased, the expression of a number of specific TFs and nuclear hormone receptors are affected. This suggests that the expression of these particular factors is regulated by DAF-16, and that they may play an important role in the up-regulation of the indirect DAF-16 targets associated with increased lifespan.

These factors are summarised in Table 5.9 and discussed in this section.

Transcript ID	Expression fold-change	Homology to	E value	P value	*C. elegans* homologue
Potential longevity-associated transcription factors					
R13H8.1b.1	1.6	*C. elegans* DAF-16	10^{-37}	0.0008	DAF-16
F26D12.1	2.4	Mouse QRF-1	2^{-29}	0.00123	FKH-7
Potential ageing-associated transcription factors					
C27C12.2.1	0.4	Mouse Krox-20	3^{-46}	0.00024	Zinc finger
ZK1067.2.1	0.3	*Drosophila* STC	2^{-24}	0	Zinc finger

Table 5.9: Likely ageing- and longevity-associated transcription factors whose expression appears to be regulated by DAF-16, indicating the possible *C. elegans* homologue.

Several DAF-16-regulated proteins show high similarity to transcription factors in Transfac. If available, annotations for these proteins were obtained from WormBase. In order to confirm that these proteins truly represent DAF-16-regulated transcription factors, they were PSI-BLASTED against the whole non-redundant Genbank. This method allowed the likely closest relatives of each protein to be identified. Also, since not all transcription factors are represented in Transfac, it increased the likelihood of identifying the true orthologues of a protein, and therefore of providing functional information about the specific family member in question, rather than simply general information concerning the function of the family. Unfortunately, however, the annotation of Genbank entries was often less extensive than that of WormBase.

Transcript R13H8.1b.1 is highly similar to both DAF-16a1 and DAF-16a2 (E-values = $7e^{-38}$), confirming the annotation of this gene. However the expression changes for this gene, while significant, are relatively low: R13H8 is up-regulated only 1.6-fold between DAF-2/DAF-16 and DAF-2 mutants (P = 0.00083). This suggests that while the ILS pathway may exert some transcriptional control on DAF-16 expression, in addition to regulation of its function by phosphorylation, this transcriptional control is relatively weak.

The strongly up-regulated transcript F26D12.1 showed similarity to the mouse glutamine-rich factor, QRF-1 (E value = $2e^{-29}$). This transcript corresponds to the *C. elegans* fork-head transcription factor FKH-7, the specific function of which is unknown

since RNAi does not produce any obvious phenotype (Hope *et al.*, 2003). FKH-7 is likely to have redundant functions with several other forkhead TFs, all of which are expressed in neurones and proposed to have a role in development of the nervous system (Li & Tucker, 1993).

In long-lived DAF-2 mutants, FKH-7 is up-regulated by 2.4-fold. This suggests that, by altering expression of its target genes, it is likely to initiate a variety of downstream responses that may contribute to longevity of these mutants. FKH-7 appears to be a direct transcriptional target of DAF-16, containing a DBE in its 5' upstream region. In order to confirm that this site is functional, and hence that FKH-7 is a true direct DAF-16 target, it will be useful to use phylogenetic footprinting to verify the presence of the DBE in orthologues in other species. It will be important to characterise this factor, its binding sequence and its target genes, in order to more fully understand its contribution to lifespan.

As shown in Figure 5.6, several FOX binding sites that are not specific for DAF-16 are over-represented in longevity-associated genes. It is possible that in DAF-2 mutants, DAF-16 stimulates expression of FKH-7, which in turn activates a set of downstream longevity-associated genes, via FOX binding sites.

In addition, two transcripts down-regulated in DAF-2 mutants show similarity to zinc-finger transcription factors. Search of Transfac using the protein sequence of C27C12.2.1 identified similarity to mouse Krox-20 (E value = $3e^{-46}$), and of ZK1067.2.1 to the Drosophila factor, shuttle craft (STC) (E value = $2e^{-24}$). PSI-BLAST identified two *C. elegans* zinc finger proteins of unknown function as their closest relatives (Genbank identifiers: 17550662 and 17538027 respectively, E values both = 0.0).

Krox-20 and STC are thought to be involved in neuronal development (Stroumbakis *et al.*, 1996; De *et al.*, 2003; Parkinson *et al.*, 2004). Both C27C12.2.1 and ZK1067.2.1 were down-regulated by approximately 0.4-fold in DAF-2 mutants compared to DAF-2/DAF-16 double mutants. This suggests that the targets of these transcription factors must be down-regulated when DAF-16 is high, because they promote ageing, or otherwise interfere with the long-lived phenotype of DAF-2 mutants. It remains unclear how Krox-20 and STC may promote ageing, but it seems likely that they will have as yet unknown effects that may contribute to control of lifespan. Consistent with the down-regulation of Krox-20 in DAF-2 mutants, where the action of insulin is blocked, Krox-20 expression is known to be stimulated by insulin (Keeton *et al.*, 2003).

The expression of approximately 26 nuclear hormone receptors is also regulated by DAF-16, although these receptors are at present all orphan receptors with no known ligand or DNA binding specificity. Characterisation of these factors will be important for fully understanding the mechanism by which many genes are indirectly regulated by

DAF-16 to cause an extension of lifespan. It is likely that one of these will be the proposed DAF-12 ligand, responsible for the non-cell-autonomous actions of DAF-16.

Further work is needed to identify the transcriptional targets of all of these TFs and nuclear hormone receptors, in order to provide further clues about the mechanism of regulation of longevity by DAF-16. At present we can conclude that it is very likely that DAF-16 modifies not just the expression of a single set of longevity- and ageing-associated genes, but also of multiple transcription factors that will in turn alter the expression of their own target genes. The lifespan of the organism is likely to be determined by the combined action of all of these genes.

5.3.4 Feedback control of the DAF-16 longevity pathway

One important direct target of DAF-16 is DAF-16 itself. When DAF-16 is active in the nucleus, its expression is increased, leading to a positive feedback loop that strengthens the increase in lifespan.

Expression of neuronally expressed insulin-like peptides is altered in DAF-2 mutants, indicating a feedback loop that may contribute to the cell non-autonomous effects of DAF-16 (Apfeld & Kenyon, 1998). However, the 5' and first intron sequences of these peptides, INS-7, INS-22 and INS-23, lack DAF-16 binding elements, suggesting that they are all indirectly regulated by DAF-16. (INS-22 does contain an over-representation of binding sites for FOXO1 and FOXO4, mammalian homologues of DAF-16, suggesting that it may be directly regulated). Interestingly, while INS-22 and INS-23 are up-regulated two-fold, INS-7 is down-regulated in DAF-2 mutants. This suggests that different insulin-like peptides may play different roles in insulin signalling and that DAF-16 expression is under both positive and negative regulatory control.

In DAF-2 RNAi treated *C. elegans*, the down-regulation of INS-7 has previously been noted by Murphy *et al.* (2003). The authors subjected wildtype *C. elegans* to INS-7 RNAi and demonstrated an increase in lifespan, consistent with the role of INS-7 as a DAF-2 agonist involved in positive feedback control of the ILS pathway. It will be interesting to use a similar method to investigate the role of INS-22 and INS-23.

Other direct and indirect targets of DAF-16 are also likely to contribute to feedback control of genes at all levels of the cascade. For example, direct or indirect targets may modify the activity or expression of DAF-16 itself, or of other DAF-16 targets. The result is a tightly controlled network of interactions, and a large number of genes which act in concert to cause major changes in gene expression and precise control of lifespan.

5.4 Discussion

Six specific aims for this work were described in Section 5.1.3 and they have all been met. The chapter first identified DAF-16 binding sites amongst DAF-16 regulated genes (Aim 1). This has permitted the DAF-16 regulated genes to be divided into direct and indirect targets (Aim 2; see Section 5.3.1.1) and enabled an investigation, via the literature, of the possible role that these targets may play in the regulation of lifespan by DAF-16 (Aim 3; Section 5.3.1.2). In addition, Section 5.3.1.4 describes identification of a number of ageing- and longevity-associated transcription factors (Aim 4). With binding sites over- or under-represented within DAF-16 targets, these factors represent possible candidates for contributing to the control of lifespan. In the case of the majority of these transcription factors, to our knowledge, a role in longevity or ageing has not previously been considered. In Section 5.3.3 DAF-16-regulated transcription factors and hormones, with the potential to provide feedback regulation of the ILS pathway, have been identified (Aim 5). Analysis of the known targets of longevity- or ageing-associated transcription factors has provided clues about additional mechanisms by which lifespan can be modified. Finally, the combined results of this chapter have enabled a clearer picture to be drawn of the mechanisms by which lifespan is controlled (Aim 6).

5.4.1 Complex structure of the DAF-16 regulatory cascade

It should be noted that, while the mechanism of control of lifespan by DAF-16 is illustrated in Figure 5.5 as a two stage process, this is not necessarily the case. Furthermore, it is very likely that the true process is a multi-stage cascade of genes and TFs, each being activated by upstream factors and in turn activating or repressing other factors downstream.

There are two main advantages for the use of such a complex network/cascade in the control of lifespan:

1. The presence of DAF-16 binding sites within the promoters of certain genes may be required for longevity would be detrimental in other situations.

2. The requirement for and interaction of a larger number of TFs is likely to provide more precise control of longevity. The presence or absence of multiple other factors (eg heat shock proteins) may act as checkpoints to determine the outcome of the pathway. In support of this, Lin *et al.* (2001) have shown that nuclear localisation of DAF-16 is insufficient for an increase in lifespan, indicating that other factors are required in addition.

Perhaps the direct DAF-16 targets are the evolutionarily most ancient determinants of longevity, but later other gene classes have been recruited as their functions became more

specialised. This would tend to lead to modularity of transcriptional control, in which multiple groups of genes, each controlled independently, were integrated by the action of another TF on their own specific regulatory factors.

The work described in this chapter has increased our understanding of this complex cascade that regulates lifespan. It has shown that insulin-like peptides and DAF-16 itself are regulated by DAF-16, providing both positive and negative feedback control of ILS. In addition, several transcription factors, FKH-7 and homologues of Krox-20 and STC, are controlled by DAF-16. These factors are likely candidates for contributing to the control of lifespan, by stimulating the expression of their own target genes. As described by Murphy *et al.* (2003), the mechanisms that extend lifespan in DAF-2 mutants are likely to be many, involving the cumulative action of numerous transcription factors and their target genes.

5.4.2 Limitations of the techniques used and minimising their impact

There are two general problems associated with the use of motif searching algorithms such as Clover, that has been used throughout this study. Firstly, these methods heavily rely on the completeness and quality of the database of known TFs that is used. Secondly, since TFBSs are relatively small, many false positives occur due to random occurrence of the same combination of bases.

The first problem, that of the lack of completeness of the databases of known transcription factors is the main limitation of the current work. The binding site of DAF-16, for example, a transcription factor central to the current work, is not found in the publicly available Transfac database and is one of only 5 nematode TFBSs in Transfac Professional v8.1. Our lack of knowledge of TFBSs is a problem that can only be alleviated with time. The current work uses the most up-to-date source of binding matrices, and represents the clearest picture that can be provided at this time.

In the future it will be important to use experimental techniques to characterise the binding specificities of more *C. elegans* transcription factors, since this species represents an important model system. It may also be valuable to establish orthology relationships between mammalian and *C. elegans* TFs, to facilitate the reliable inheriting of functional information for factors where the *C. elegans* homologue is yet to be identified. Characterisation of the interactions between factors that may modify binding or regulatory activity will also be important. Work towards these goals will greatly increase the information that can be gained using this method about the regulation of many physiological processes, including ageing.

The second problem, that of false positives, derives from the fact that TFs rely upon additional signals, such as interaction with other TFs and variations in chromatin structure, in combination with the presence of their binding consensus sequence. Clover was selected because it is able to counter this problem relatively effectively. This is achieved by detecting a *relative* enrichment of a certain site in a group of functionally related or co-regulated sequences is detected, compared to a set of background sequences. Provided that both sets of sequences are derived from the same genome, the false positive rate would be expected to be constant between them and any difference in occurrence is likely to represent functional sites.

Whilst our knowledge of the characteristics of TFBSs alone is insufficient to drive their identification, this information is valuable in guiding other techniques to select the correct motifs from a pool of candidates that match a binding matrix. Many motif searching algorithms do not exploit this useful biological knowledge. For example, motif searching algorithms generally do not take into account phylogenetic information, and often have no requirement for the identified motifs to be found in the same order or orientation in different genes. Improvements in accuracy are being made as biological information is incorporated into models. For example, regulatory regions often contain multiple, overlapping copies of the same TFBS (Papatsenko *et al.*, 2002). Clover is the first method to make use of this information, by combining the individual score matrices for each site to generate the overall occupancy score for the regulatory region. In order to increase the accuracy of such methods, a priority for future work must be improving models of TFBS occupancy, by determining the effects of repeated and overlapping sites on TF binding.

The methods used throughout this work have been designed to minimise the effect of these problems by:

- Limiting the sequences searched to those most likely to contain functional TFBSs, by excluding coding sequences and regions more than 1kb from the transcription start site.
- Looking for over/under-representation of TFBSs with respect to an appropriate background sequence set, rather than simply the presence/absence of certain motifs.
- Only considering TFBSs with high probabilities of being truly over/under-represented ($P<0.01$).
- Using the advanced statistics of Clover to incorporate biological knowledge by giving additional weight to repeated and overlapping binding sites

It is hoped that by the use of these techniques the ageing- and longevity-associated TFBSs identified here have enabled a clearer understanding of the mechanisms of lifespan control.

5.4.3 Final conclusions

It must be remembered that conserved or over-represented motifs identified by phylogenetic footprinting or motif searching may represent sites involved in functions other than binding of TFs. For example, they may be required to confer mRNA stability (Roth *et al.*, 1998). Further analysis is required before it can be concluded that the identified motifs are true TFBSs. The use of several complementary computational techniques, e.g. motif searching and phylogenetic footprinting, may be of use in this respect, but often experimental confirmation will be necessary. Important future work will involve using phylogenetic footprinting to confirm the conservation of interesting sites identified in this study in related species. In addition it will be necessary to investigate experimentally the possible role of the transcription factors identified here in ageing. Null mutations or RNAi can be used to determine the effect that the protein of interest has on lifespan. Comparison of the data obtained here to that for another species would also be valuable. Considerable work will be required before we have a complete understanding of the control of lifespan.

The implications of this work extend beyond the control of *C. elegans* lifespan, since mutations in the ILS pathway have been shown to extend lifespan in an evolutionarily diverse set of organisms, including mammals (Bluher *et al.*, 2003; Holzenberger *et al.*, 2003). Hence, it is possible that the mechanisms identified here are involved in the regulation of human ageing. This work therefore has implications for the eventual design of therapeutic interventions able to slow ageing and ageing-associated disease.

Chapter 6

Conclusions

6.1 Information gained concerning uncoupling protein and membrane protein structure

The aim of this thesis, stated in the Introduction (Chapter 1), was to increase our understanding of the mechanisms of ageing. In some chapters this has been achieved, whereas in others the results contribute mainly to our understanding of membrane protein structure and prediction. This final chapter attempts to place these findings in terms of the wider field, and to identify important areas for future work.

In Chapter 2, modelling of uncoupling protein structure was attempted using experimental data from the literature. Now that the structure of a homologue of the UCPs has been solved, the value of this work lies mainly in determining the potential of such a method for modelling of other TM proteins. Unfortunately, however, most models were supported by the majority of the pieces of experimental data, making it impossible to definitively either support or refute any of them. The problem with the method lies in the difficulty in interpreting the results from mutational experiments: it is often impossible to determine whether the loss of function that occurs on mutation is caused directly by the loss of a participating residue, or indirectly due to disruption of the native conformation. Hence, much of the experimental evidence described here is not definitive enough to use for model prediction in this context. These results suggest that this technique is unlikely to be effective for modelling of other proteins unless a very large amount of reliable mutagenesis data is available. It seems that the main value of mutagenesis data in modelling will be in providing additional support for models proposed by other methods.

Chapter 4 built upon the analysis of Chapter 2, using a predictive method based on the general principles of TM protein structure. It was felt that in order to maximise predictive accuracy, such a method required the greatest possible understanding of these

structural features. To this end a detailed analysis of the currently available polytopic TM protein structures were performed in Chapter 3. 24 TM protein families were identified that were represented in the Protein Data Bank (as of January 2004). This number is more than twice that available when the last comprehensive analysis was performed in 2001 (Ulmschneider & Sansom, 2001). It will be important for this analysis to be updated regularly in the future, as the number of available structures increases further.

Basic analysis of these TM protein structures were performed, generally confirming the results of previous, smaller studies by identifying the preferences of different residues for different TM environments. The results clearly show that the majority of TM helix-helix contacts are made by either small relatively polar residues, particularly glycine, alanine and serine, or large hydrophobic residues. Many of the charged and aromatic residues show strong preferences for lipid-tail-accessibility.

The work described in Chapter 4 indicates that some charged residues in the TM lipid-tail region show a preference for lipid-tail-accessible not buried positions. Contradictory results on this point had been obtained in previous analyses (Javadpour *et al.*, 1999; Ulmschneider & Sansom, 2001). The finding was unexpected and possible explanations for the presence of lipid-tail-accessible charged residues were investigated for the first time. The results show that the majority of lipid-tail-accessible charged residues are not paired, but do satisfy their hydrogen bonding potential in other ways, namely by interaction with the head-group region. Interestingly, almost one third of the interactions with other residues are intrahelical. As described in Chapter 4, further experimental work is required to determine the effect of lipid-tail-accessible charged and polar residues and of intrahelical hydrogen bonds on TM protein stability and function.

The work also showed that the preferences of residues for buried or accessible positions in TM proteins cannot simply be predicted by the use of a traditional hydrophobicity scale. As an alternative, a 'lipid-tail-accessibility scale' was developed that represents the residue preferences that are found in TM proteins. As a result of this work we now have a much clearer and more confident understanding of helix-helix packing in TM proteins.

The LA scale, together with other knowledge gained during the analysis of Chapter 3, was applied to the prediction of UCP structure in Chapter 4. These findings are also applicable to the modelling of other TM proteins lacking structural information. Chapter 4 first investigated the use of these parameters for prediction of buried and accessible residues, using proteins of known 3D structure. The resulting method was then used to score the predefined models of UCP structure from Chapter 2, in an attempt to select the most likely model.

The value of the approach taken to TM protein modelling in this work lies in the potential to produce models in the absence of any structural information for the family

concerned, and without relying upon the use of structural information from water-soluble proteins. The method was relatively good at detecting buried helix faces in TM proteins of known structure and has enabled the UCP models to be ranked in order of likelihood. However, the model proposed for the UCPs is highly tentative. The model selected did not score significantly higher than other models by any one method used here. Comparison of the model with the known structure of a UCP homologue, solved after the modelling was completed, has enabled the reasons for this to be determined.

The main problems with the approach seem to be that the method models TM helices as ideal helices arranged in parallel and the suggestion that the protein was a dimer. Large deviations from these assumptions, such as the kinked, highly tilted or partially-spanning helices seen in the actual structure, caused inaccuracies in the modelling. Hence, the method is unsuitable for model scoring for proteins which are likely to have kinked helices due to transmembrane proline residues. In addition, the method may be a little less useful for the prediction of models for families for which there are no simplifying symmetry constraints, leading to a huge number of potential models that must be considered. In these cases, however, the method remains able to provide valuable structural information, by identifying the buried faces of TM helices. Such information will be particularly useful in generating hypotheses of helix packing that can be tested experimentally using cross-linking or mutagenesis techniques.

The work in Chapter 4 has involved in investigation into not only UCP structure, but also into our current understanding of membrane protein structure and our ability to model it. The work has shown that, whilst the model proposed for the UCPs showed similarities with the correct structure, the technique is too simple and not yet good enough to predict models with confidence. More membrane protein structures will need to be solved, and our understanding of the ways in which they pack to obtain stability will need to increase, before this is possible. In addition, reliable information about the oligomeric state of the protein is essential to translating the predictions of buried helix faces into possible arrangements of TM helices for scoring.

Were this work to be repeated, given our current knowledge, the modelling would obviously be approached somewhat differently. More complex modelling procedures, in which helix tilting and kinking are considered, would be needed to increase predictive accuracy. It may also be valuable to use a higher resolution approach, incorporating molecular factors such as surface shape and charge complementarity, or to consider known helix-helix packing motifs.

It is clear from this work that the challenge of predicting their tertiary structure from sequence has yet to be solved. However, the structures of membrane proteins are limited by their unusual environment and are less diverse than water-soluble proteins. As a

result it is likely that, were equal resources devoted to the study of TM and water-soluble proteins, the prediction of structure from sequence would be achieved first in TM proteins.

The results presented in this part of the thesis will be of use in the long term in the understanding of TM protein structure and prediction of TM helix packing. At present it appears that considerably more TM protein structures will be needed before these goals are achieved. In addition to a lack of data concerning TM helix packing mechanisms, identification of protein-protein interfaces and determination of oligomeric state from monomeric structure remain major pitfalls in structure prediction methods such as those developed in Chapter 4. However, with the launch of the first structural genomics program focused entirely on TM proteins in 2003 (Kyogoku *et al.*, 2003), our knowledge of membrane protein structure is likely to increase substantially in the next few years.

6.2 Information gained concerning the mechanism of lifespan regulation by DAF-16

Chapter 5 has identified direct and indirect targets of DAF-16 and investigated, via the literature, the possible role that these targets may play in the regulation of lifespan by DAF-16. This has enabled a clearer picture to be drawn of the mechanisms by which lifespan is controlled. In addition, the results of this work have included the identification of a number of ageing- and longevity-associated transcription factors. With binding sites over- or under-represented within DAF-16 targets, these factors represent possible candidates for contributing to the control of lifespan. In the case of the majority of these transcription factors, to my knowledge, a role in longevity or ageing has not previously been considered. Many of the factors are known to be involved in the embryonic development of particular tissues, and a role in the adult has not been identified. This work suggests that they may also be involved in the regulation of stress responses or other pathways in the corresponding adult tissue, as has been demonstrated for Nkx2.5. Analysis of the known targets of longevity- or ageing-associated transcription factors has provided clues about additional mechanisms by which lifespan can be modified.

This work has also highlighted the complex structure of the DAF-16 regulatory cascade. It has shown that insulin-like peptides and DAF-16 itself are regulated by DAF-16, providing both positive and negative feedback control of ILS. In addition, several transcription factors, FKH-7 and homologues of Krox-20 and STC, are controlled by DAF-16. These factors are likely candidates for contributing to the control of lifespan, by stimulating the expression of their own target genes.

Important future work will involve using phylogenetic footprinting to confirm the conservation of interesting sites identified in this study in related species. It will also to necessary to identify and characterise the *C. elegans* factors that bind to the over-represented mammalian sites identified. Null mutations or RNAi could be used to determine the effect that the transcription factor of interest has on lifespan. Considerable work will be required before we have a complete understanding of the control of lifespan.

6.3 Final conclusions

The implications of this work extend beyond the study of *C. elegans* lifespan. Mutations in the ILS pathway have been shown to extend lifespan in a wide range of organisms, including mammals (Bluher *et al.*, 2003; Holzenberger *et al.*, 2003). Similarly, the uncoupling proteins are present in virtually all species, including humans. Hence it is possible that the mechanisms identified and studied here are involved in human ageing. Such knowledge may eventually lead to the development of therapeutic interventions to slow ageing or control progression of ageing-associated diseases.

Bibliography

Abramson, J., Riistama, S., Larsson, G., Jasaitis, A., Svensson-Ek, M., Laakkonen, L., Puustinen, A., Iwata, S. & Wikstrom, M. (2000). The structure of the ubiquinol oxidase from Escherichia coli and its ubiquinone binding site. *Nat Struct Biol*, **7**, 910–7.

Abramson, J., Smirnova, I., Kasho, V., Verner, G., Kaback, H. & Iwata, S. (2003). Structure and mechanism of the lactose permease of Escherichia coli. *Science*, **301**, 610–5.

Adamian, L. & Liang, J. (2001). Helix-helix packing and interfacial pairwise interactions of residues in membrane proteins. *J Mol Biol*, **311**, 891–907.

Adamian, L. & Liang, J. (2002). Interhelical hydrogen bonds and spatial motifs in membrane proteins: polar clamps and serine zippers. *Proteins*, **47**, 209–18.

Adamian, L. & Liang, J. (2003). Interhelical hydrogen bonds in transmembrane region are important for function and stability of Ca -transporting ATPase. *Cell Biochem Biophys*, **39**, 1–12.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. *J Mol Biol*, **215**, 403–10.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389–402.

Alvarez, R., de Andres, J., Yubero, P., Vinas, O., Mampel, T., Iglesias, R., Giralt, M. & Villarroya, F. (1995). A novel regulatory pathway of brown fat thermogenesis. Retinoic acid is a transcriptional activator of the mitochondrial uncoupling protein gene. *J Biol Chem*, **270**, 5666–73.

Amati, B., Brooks, M., Levy, N., Littlewood, T., Evan, G. & Land, H. (1993). Oncogenic activity of the c-Myc protein requires dimerization with Max. *Cell*, **72**, 233–45.

An, J. & Blackwell, T. (2003). SKN-1 links C. elegans mesendodermal specification to a conserved oxidative stress response. *Genes Dev*, **17**, 1882–93.

Antier, D., Carswell, H., Brosnan, M., Hamilton, C., Macrae, I., Groves, S., Jardine, E., Reid, J. & Dominiczak, A. (2004). Increased levels of superoxide in brains from old female rats. *Free Radic Res*, **38**, 177–83.

Apfeld, J. & Kenyon, C. (1998). Cell nonautonomy of C. elegans daf-2 function in the regulation of diapause and life span. *Cell*, **95**, 199–210.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N.J., Oinn, T.M., Pagni, M., Servant, F., Sigrist, C.J. & Zdobnov, E.M. (2000). InterPro–an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics*, **16**, 1145–50.

Aquila, H., Link, T.A. & Klingenberg, M. (1985). The uncoupling protein from brown fat mitochondria is related to the mitochondrial ADP/ATP carrier. Analysis of sequence homologies and of folding of the protein in the membrane. *EMBO J*, **4**, 2369–76.

Arechaga, I., Raimbault, S., Prieto, S., Levi-Meyrueis, C., Zaragoza, P., Miroux, B., Ricquier, D., Bouillaud, F. & Rial, E. (1993). Cysteine residues are not essential for uncoupling protein function. *Biochem J*, **296**, 693–700.

Arechaga, I., Ledesma, A. & Rial, E. (2001). The mitochondrial uncoupling protein UCP1: a gated pore. *IUBMB Life*, **52**, 165–73.

Arkin, I., Brunger, A. & Engelman, D. (1997). Are there dominant membrane protein families with a given number of helices? *Proteins*, **28**, 465–6.

Armstrong, M.B. & Towle, H.C. (2001). Polyunsaturated fatty acids stimulate hepatic UCP-2 expression via a PPARalpha-mediated pathway. *Am J Physiol Endocrinol Metab*, **281**, E1197–E1204.

Arselin, G., Giraud, M., Dautant, A., Vaillier, J., Brethes, D., Coulary-Salin, B., Schaeffer, J. & Velours, J. (2003). The GxxxG motif of the transmembrane domain of subunit e is involved in the dimerization/oligomerization of the yeast ATP synthase complex in the mitochondrial membrane. *Eur J Biochem*, **270**, 1875–84.

Arsenijevic, D., Onuma, H., Pecqueur, C., Raimbault, S., Manning, B.S., Miroux, B., Couplan, E., Alves-Guerra, M.C., Goubern, M., Surwit, R., Bouillaud, F., Richard,

D., Collins, S. & Ricquier, D. (2000). Disruption of the uncoupling protein-2 gene in mice reveals a role in immunity and reactive oxygen species production. *Nat Genet*, **26**, 435–9.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25–9.

Baes, M. & Declercq, P. (1998). Gel-shift analysis and identification of RXREs and RAREs by PCR-based selection. *Methods Mol Biol*, **89**, 377–88.

Bailey, T. & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, **2**, 28–36.

Bairoch, A. & Apweiler, R. (1997). The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res*, **25**, 31–6.

Baker, E. & Hubbard, R. (1984). Hydrogen bonding in globular proteins. *Prog Biophys Mol Biol*, **44**, 97–179.

Barazzoni, R. & Nair, K.S. (2001). Changes in uncoupling protein-2 and -3 expression in aging rat skeletal muscle, liver, and heart. *Am J Physiol Endocrinol Metab*, **280**, E413–9.

Barbier, O., Villeneuve, L., Bocher, V., Fontaine, C., Torra, I., Duhem, C., Kosykh, V., Fruchart, J., Guillemette, C. & Staels, B. (2003). The UDP-glucuronosyltransferase 1A9 enzyme is a peroxisome proliferator-activated receptor alpha and gamma target gene. *J Biol Chem*, **278**, 13975–83.

Barbieri, M., Bonafe, M., Rizzo, M., Ragno, E., Olivieri, F., Marchegiani, F., Franceschi, C. & Paolisso, G. (2004). Gender specific association of genetic variation in peroxisome proliferator-activated receptor (PPAR)gamma-2 with longevity. *Exp Gerontol*, **39**, 1095–100.

Bartlett, G.J., Porter, C.T., Borkakoti, N. & Thornton, J.M. (2002). Analysis of catalytic residues in enzyme active sites. *J Mol Biol*, **324**, 105–21.

Bass, R.B., Strop, P., Barclay, M. & Rees, D.C. (2002). Crystal structure of Escherichia coli MscS, a voltage-modulated and mechanosensitive channel. *Science*, **298**, 1582–7.

Belrhali, H., Nollert, P., Royant, A., Menzel, C., Rosenbusch, J., Landau, E. & Pebay-Peyroula, E. (1999). Protein, lipid and water organization in bacteriorhodopsin crystals: a molecular view of the purple membrane at 1.9 A resolution. *Structure Fold Des*, **7**, 909–17.

Berezikov, E., Guryev, V., Plasterk, R. & Cuppen, E. (2004). CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. *Genome Res*, **14**, 170–8.

Bernstein, F., Koetzle, T., Williams, G., Meyer, J.r., Brice, M., Rodgers, J., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol*, **112**, 535–42.

Bertero, M., Rothery, R., Palak, M., Hou, C., Lim, D., Blasco, F., Weiner, J. & Strynadka, N. (2003). Insights into the respiratory electron transfer pathway from the structure of nitrate reductase A. *Nat Struct Biol*, **10**, 681–7.

Bertolino, E., Reimund, B., Wildt-Perinic, D. & Clerc, R. (1995). A novel homeobox protein which recognizes a TGT core and functionally interferes with a retinoid-responsive motif. *J Biol Chem*, **270**, 31178–88.

Bezaire, V., Hofmann, W., Kramer, J.K., Kozak, L.P. & Harper, M.E. (2001). Effects of fasting on muscle mitochondrial energetics and fatty acid metabolism in Ucp3(-/-) and wild-type mice. *Am J Physiol Endocrinol Metab*, **281**, E975–82.

Bienengraeber, M., Echtay, K.S. & Klingenberg, M. (1998). H+ transport by uncoupling protein (UCP-1) is dependent on a histidine pair, absent in UCP-2 and UCP-3. *Biochemistry*, **37**, 3–8.

Bigelow, H., Wenick, A., Wong, A. & Hobert, O. (2004). CisOrtho: a program pipeline for genome-wide identification of transcription factor target genes using phylogenetic footprinting. *BMC Bioinformatics*, **5**, 27–34.

Bisaccia, F., Zara, V., Capobianco, L., Iacobazzi, V., Mazzeo, M. & Palmieri, F. (1996). The formation of a disulfide cross-link between the two subunits demonstrates the dimeric structure of the mitochondrial oxoglutarate carrier. *Biochim Biophys Acta*, **1292**, 281–88.

Blanchette, M. & Tompa, M. (2002). Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res*, **12**, 739–48.

Blanchette, M. & Tompa, M. (2003). FootPrinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res*, **31**, 3840–2.

Blanchette, M., Schwikowski, B. & Tompa, M. (2002). Algorithms for phylogenetic footprinting. *J Comput Biol*, **9**, 211–23.

Bluher, M., Kahn, B. & Kahn, C. (2003). Extended longevity in mice lacking the insulin receptor in adipose tissue. *Science*, **299**, 572–4.

Boone, A. & Vijayan, M. (2002). Glucocorticoid-mediated attenuation of the hsp70 response in trout hepatocytes involves the proteasome. *Am J Physiol Regul Integr Comp Physiol*, **283**, R680–7.

Bouillaud, F., Weissenbach, J. & Ricquier, D. (1986). Complete cDNA-derived amino acid sequence of rat brown fat uncoupling protein. *J Biol Chem*, **261**, 1487–90.

Bouillaud, F., Arechaga, I., Petit, P.X., Raimbault, S., Levi-Meyrueis, C., Casteilla, L., Laurent, M., Rial, E. & Ricquier, D. (1994). A sequence related to a DNA recognition element is essential for the inhibition by nucleotides of proton transport through the mitochondrial uncoupling protein. *EMBO J*, **13**, 1990–7.

Bouillaud, F., Couplan, E., Pecqueur, C. & Ricquier, D. (2001). Homologues of the uncoupling protein from brown adipose tissue (UCP1): UCP2, UCP3, BMCP1 and UCP4. *Biochim Biophys Acta*, **1504**, 107–19.

Bowie, J.U. (1997). Helix packing in membrane proteins. *J Mol Biol*, **272**, 780–9.

Bowser, R., Giambrone, A. & Davies, P. (1995). FAC1, a novel gene identified with the monoclonal antibody Alz50, is developmentally regulated in human brain. *Dev Neurosci*, **17**, 20–37.

Branco, M., Ribeiro, M., Negrao, N. & Bianco, A.C. (1999). 3,5,3'-Triiodothyronine actively stimulates UCP in brown fat under minimal sympathetic activity. *Am J Physiol*, **276**, E179–87.

Brandolin, G., Dupont, Y. & Vignais, P.V. (1982). Exploration of the nucleotide binding sites of the isolated ADP/ATP carrier protein from beef heart mitochondria. 2. Probing of the nucleotide sites by formycin triphosphate, a fluorescent transportable analogue of ATP. *Biochemistry*, **21**, 6348–53.

Bretscher, M.S. & Munro, S. (1993). Cholesterol and the Golgi apparatus. *Science*, **261**, 1280–1.

Brogiolo, W., Stocker, H., Ikeya, T., Rintelen, F., Fernandez, R. & Hafen, E. (2001). An evolutionarily conserved function of the Drosophila insulin receptor and insulin-like peptides in growth control. *Curr Biol*, **11**, 213–21.

Busuttil, R., Dolle, M., Campisi, J. & Vijga, J. (2004). Genomic instability, aging, and cellular senescence. *Ann N Y Acad Sci*, **1019**, 245–55.

Bywater, R.P., Thomas, D. & Vriend, G. (2001). A sequence and structural study of transmembrane helices. *J Comput Aided Mol Des*, **15**, 533–52.

Cadenas, S., Echtay, K.S., Harper, J.A., Jekabsons, M.B., Buckingham, J.A., Grau, E., Abuin, A., Chapman, H., Clapham, J.C. & Brand, M.D. (2002). The basal proton conductance of skeletal muscle mitochondria from transgenic mice overexpressing or lacking uncoupling protein-3. *J Biol Chem*, **277**, 2773–8.

Campagne, F. & Weinstein, H. (1999). Schematic representation of residue-based protein context-dependent data: an application to transmembrane proteins. *J Mol Graph Model*, **17**, 207–13.

Casteilla, L., Rigoulet, M. & Penicaud, L. (2001). Mitochondrial ROS metabolism: modulation by uncoupling proteins. *IUBMB Life*, **52**, 181–8.

Cha, S.H., Fukushima, A., Sakuma, K. & Kagawa, Y. (2001). Chronic docosahexaenoic acid intake enhances expression of the gene for uncoupling protein 3 and affects pleiotropic mRNA levels in skeletal muscle of aged C57BL/6NJcl mice. *J Nutr*, **131**, 2636–42.

Chang, C.H., el Kabbani, O., Tiede, D., Norris, J. & Schiffer, M. (1991). Structure of the membrane-bound protein photosynthetic reaction center from Rhodobacter sphaeroides. *Biochemistry*, **30**, 5352–60.

Chang, D.K., Cheng, S.F., Trivedi, V.D. & Lin, K.L. (1999). Proline affects oligomerization of a coiled coil by inducing a kink in a long helix. *J Struct Biol*, **128**, 270–9.

Chang, G. (2003). Structure of MsbA from Vibrio cholera: a multidrug resistance ABC transporter homolog in a closed conformation. *J Mol Biol*, **330**, 419–30.

Chang, G. & Roth, C. (2001). Structure of MsbA from E. coli: a homolog of the multidrug resistance ATP binding cassette (ABC) transporters. *Science*, **293**, 1793–800.

Chang, G., Spencer, R.H., Lee, A.T., Barclay, M.T. & Rees, D.C. (1998). Structure of the MscL homolog from Mycobacterium tuberculosis: a gated mechanosensitive ion channel. *Science*, **282**, 2220–6.

Chapman, T. & Partridge, L. (1996). Female fitness in Drosophila melanogaster: an interaction between the effect of nutrition and of encounter rate with males. *Proc R Soc Lond B Biol Sci*, **263**, 755–9.

Chapman, T., Miyatake, T., Smith, H. & Partridge, L. (1998). Interactions of mating, egg production and death rates in females of the Mediterranean fruit fly, Ceratitis capitata. *Proc R Soc Lond B Biol Sci*, **265**, 1879–94.

Chen, C.M. & Chen, C.C. (2003). Computer simulations of membrane protein folding: structure and dynamics. *Biophys J*, **84**, 1902–8.

Chen, Q., Hertz, G. & Stormo, G. (1995). MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput Appl Biosci*, **11**, 563–6.

Chi, N. & Epstein, J. (2002). Getting your Pax straight: Pax proteins in development and disease. *Trends Genet*, **18**, 41–7.

Chin, C.N. & von Heijne, G. (2000). Charge pair interactions in a model transmembrane helix in the ER membrane. *J Mol Biol*, **303**, 1–5.

Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. *J Mol Biol*, **105**, 1–12.

Clancy, D.J., Gems, D., Harshman, L.G., Oldham, S., Stocker, H., Hafen, E., Leevers, S.J. & Partridge, L. (2001). Extension of life-span by loss of CHICO, a Drosophila insulin receptor substrate protein. *Science*, **292**, 104–6.

Clapham, J.C., Arch, J.R., Chapman, H., Haynes, A., Lister, C., Moore, G.B., Piercy, V., Carter, S.A., Lehner, I., Smith, S.A., Beeley, L.J., Godden, R.J., Herrity, N., Skehel, M., Changani, K.K., Hockings, P.D., Reid, D.G., Squires, S.M., Hatcher, J., Trail, B., Latcham, J., Rastan, S., Harper, A.J., Cadenas, S., Buckingham, J.A., Brand, M.D. & Abuin, A. (2000). Mice overexpressing human uncoupling protein-3 in skeletal muscle are hyperphagic and lean. *Nature*, **406**, 415–8.

Clapham, J.C., Coulthard, V.H. & Moore, G.B. (2001). Concordant mrna expression of ucp-3, but not ucp-2, with mitochondrial thioesterase-1 in brown adipose tissue and skeletal muscle in db/db diabetic mice. *Biochem Biophys Res Commun*, **287**, 1058–62.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B. & Johnston, M. (2003). Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science*, **301**, 71–6.

Conlon, E., Liu, X., Lieb, J. & Liu, J. (2003). Integrating regulatory motif discovery and genome-wide expression analysis. *Proc Natl Acad Sci (U S A)*, **100**, 3339–44.

Conroy, M.J., Westerhuis, W.H., Parkes-Loach, P.S., Loach, P.A., Hunter, C.N. & Williamson, M.P. (2000). The solution structure of Rhodobacter sphaeroides LH1beta reveals two helical domains separated by a more flexible region: structural consequences for the LH1 complex. *J Mol Biol*, **298**, 83–94.

Corpet F (1988). Multiple sequence alignment with hierarchical clustering. *Nucl. Acids Res.*, **16**, 10881–10890.

Crick, F.H.C. (1953). The packing of alpha-helices: Simple coiled-coils. *Acta Crystallog*, **6**, 689–697.

Dalgaard, L.T. & Pedersen, O. (2001). Uncoupling proteins: functional characteristics and role in the pathogenesis of obesity and Type II diabetes. *Diabetologia*, **44**, 946–65.

Dalgaard, L.T., Hansen, T., Urhammer, S.A., Drivsholm, T., Borch-Johnsen, K. & Pedersen, O. (2001). The uncoupling protein 3 55C-T variant is not associated with Type II diabetes mellitus in Danish subjects. *Diabetologia*, **44**, 1065–7.

Dayhoff, M. (1978). Matrices for detecting distant relationships. *Atlas Protein Seq Struct*, **5**, 353–358.

De, S., Shuler, C. & Turman, J.r. (2003). The ontogeny of Krox-20 expression in brainstem and cerebellar neurons. *J Chem Neuroanat*, **25**, 213–26.

de Grey, A., Baynes, J.W., Berd, D., Heward, C.B., Pawelec, G. & Stock, G. (2002). Is human ageing still mysterious enough to be left only to scientists? *BioEssays*, **24**, 667–676.

de Groot, B.L., Engel, A. & Grubmuller, H. (2001). A refined structure of human aquaporin-1. *FEBS Lett*, **504**, 206–11.

De Planque, M., Rijkers, D., Liskamp, R. & Separovic, F. (2004). The alphaM1 transmembrane segment of the nicotinic acetylcholine receptor interacts strongly with model membranes. *Magn Reson Chem*, **42**, 148–54.

de Planque, M.R., Kruijtzer, J.A., Liskamp, R.M., Marsh, D., Greathouse, D.V., Koeppe, R.E., de Kruijff, B. & Killian, J.A. (1999). Different membrane anchoring positions of tryptophan and lysine in synthetic transmembrane alpha-helical peptides. *J Biol Chem*, **274**, 20839–46.

Deisenhofer, J., Epp, O., Miki, K., Huber, R. & Michel, H. (1985). Structure of the protein subunits in the photosynthetic reaction centre of Rhodospeudomonas viridis at 3 å resolution. *Nature*, **318**, 618–624.

Deplanque, D. (2004). Cell protection through PPAR nuclear receptor activation. *Therapie*, **59**, 25–9.

Dermitzakis, E. & Clark, A. (2002). Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol*, **19**, 1114–21.

Diamond, D., Parsian, A., Hunt, C., Lofgren, S., Spitz, D., Goswami, P. & Gius, D. (1999). Redox factor-1 (Ref-1) mediates the activation of AP-1 in HeLa and NIH 3T3 cells in response to heat shock. *J Biol Chem*, **274**, 16959–64.

DiDomenico, B., Bugaisky, G. & Lindquist, S. (1982). The heat shock response is self-regulated at both the transcriptional and posttranscriptional levels. *Cell*, **31**, 593–603.

Donnelly, D., Overington, J.P., Ruffle, S.V., Nugent, J.H. & Blundell, T.L. (1993). Modeling alpha-helical transmembrane domains: the calculation and use of substitution tables for lipid-facing residues. *Protein Sci*, **2**, 55–70.

Donohoe, M., Zhang, X., McGinnis, L., Biggers, J., Li, E. & Shi, Y. (1999). Targeted disruption of mouse Yin Yang 1 transcription factor results in peri-implantation lethality. *Mol Cell Biol*, **19**, 7237–44.

Doyle, D.A., Morais Cabral, J., Pfuetzner, R.A., Kuo, A., Gulbis, J.M., Cohen, S.L., Chait, B.T. & MacKinnon, R. (1998). The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science*, **280**, 69–77.

Drummond-Barbosa, D. & Spradling, A.C. (2001). Stem sells and their progeny respond to nutritional changes during Drosophila oogenesis. *Dev Biol*, **231**, 265-278.

Dupont, Y., Brandolin, G. & Vignais, P.V. (1982). Exploration of the nucleotide binding sites of the isolated ADP/ATP carrier protein from beef heart mitochondria. 1. Probing of the nucleotide sites by naphthoyl-ATP, a fluorescent nontransportable analogue of ATP. *Biochemistry*, **21**, 6343–7.

Duret, L. & Bucher, P. (1997). Searching for regulatory elements in human noncoding sequences. *Curr Opin Struct Biol*, **7**, 399–406.

Dutzler, R., Campbell, E.B., Cadene, M., Chait, B.T. & MacKinnon, R. (2002). X-ray structure of a ClC chloride channel at 3.0 A reveals the molecular basis of anion selectivity. *Nature*, **415**, 287–94.

Dutzler, R., Campbell, E. & MacKinnon, R. (2003). Gating the selectivity filter in ClC chloride channels. *Science*, **300**, 108–12.

Echtay, K.S., Bienengraeber, M. & Klingenberg, M. (1997). Mutagenesis of the uncoupling protein of brown adipose tissue. Neutralization of E190 largely abolishes pH control of nucleotide binding. *Biochemistry*, **36**, 8253–60.

Echtay, K.S., Bienengraeber, M., Winkler, E. & Klingenberg, M. (1998). In the uncoupling protein (UCP-1) His-214 is involved in the regulation of purine nucleoside triphosphate but not diphosphate binding. *J Biol Chem*, **273**, 24368–74.

Echtay, K.S., Winkler, E., Bienengraeber, M. & Klingenberg, M. (2000a). Site-directed mutagenesis identifies residues in uncoupling protein (UCP1) involved in three different functions. *Biochemistry*, **39**, 3311–7.

Echtay, K.S., Winkler, E. & Klingenberg, M. (2000b). Coenzyme Q is an obligatory cofactor for uncoupling protein function. *Nature*, **408**, 609–13.

Echtay, K.S., Bienengraeber, M. & Klingenberg, M. (2001a). Role of intrahelical arginine residues in functional properties of uncoupling protein (UCP1). *Biochemistry*, **40**, 5243–8.

Echtay, K.S., Winkler, E., Frischmuth, K. & Klingenberg, M. (2001b). Uncoupling proteins 2 and 3 are highly active H(+) transporters and highly nucleotide sensitive when activated by coenzyme Q (ubiquinone). *Proc Natl Acad Sci (U S A)*, **98**, 1416–21.

Echtay, K.S., Roussel, D., St-Pierre, J., Jekabsons, M.B., Cadenas, S., Stuart, J.A., Harper, J.A., Roebuck, S.J., Morrison, A., Pickering, S., Clapham, J.C. & Brand, M.D. (2002). Superoxide activates mitochondrial uncoupling proteins. *Nature*, **415**, 96–9.

Eckerskorn, C. & Klingenberg, M. (1987). In the uncoupling protein from brown adipose tissue the C-terminus protrudes to the c-side of the membrane as shown by tryptic cleavage. *FEBS Lett*, **226**, 166–70.

Edman, K., Nollert, P., Royant, A., Belrhali, H., Pebay-Peyroula, E., Hajdu, J., Neutze, R. & Landau, E. (1999). High-resolution X-ray structure of an early intermediate in the bacteriorhodopsin photocycle. *Nature*, **401**, 822–6.

Eilers, M., Shekar, S.C., Shieh, T., Smith, S.O. & Fleming, P.J. (2000). Internal packing of helical membrane proteins. *Proc Natl Acad Sci (U S A)*, **97**, 5796–801.

Eilers, M., Patel, A.B., Liu, W. & Smith, S.O. (2002). Comparison of helix interactions in membrane and soluble alpha-bundle proteins. *Biophys J*, **82**, 2720–36.

Eisenberg, D., Weiss, R., Terwilliger, T. & Wilcox, W. (1982a). Hydrophobic moments and protein structure. *Faraday Symp. Chem. Soc.*, **17**, 109–120.

Eisenberg, D., Weiss, R.M. & Terwilliger, T.C. (1982b). The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature*, **299**, 371–4.

Engelman, D. & Zaccai, G. (1980). Bacteriorhodopsin is an inside-out protein. *Proc Natl Acad Sci (U S A)*, **77**, 5894–8.

Engelman, D., Steitz, T. & Goldman, A. (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Biophys Chem*, **15**, 321–53.

Essen, L., Siegert, R., Lehmann, W. & Oesterhelt, D. (1998). Lipid patches in membrane protein oligomers: crystal structure of the bacteriorhodopsin-lipid complex. *Proc Natl Acad Sci U S A*, **95**, 11673–8.

Evan, G. & Littlewood, T. (1993). The role of c-myc in cell growth. *Curr Opin Genet Dev*, **3**, 44–9.

Farrar, R.P., Martin, T.P. & Ardies, C.M. (1981). The interaction of aging and endurance exercise upon the mitochondrial function of skeletal muscle. *J Gerontol*, **36**, 642–7.

Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package) version 3.5c. *Distributed by the author. Department of Genetics, University of Washington, Seattle.*

Fleishman, S.J. & Ben-Tal, N. (2002). A novel scoring function for predicting the conformations of tightly packed pairs of transmembrane alpha-helices. *J Mol Biol*, **321**, 363–78.

Fleury, C., Neverova, M., Collins, S., Raimbault, S., Champigny, O., Levi-Meyrueis, C., Bouillaud, F., Seldin, M.F., Surwit, R.S., Ricquier, D. & Warden, C.H. (1997). Uncoupling protein-2: a novel gene linked to obesity and hyperinsulinemia. *Nat Genet*, **15**, 269–72.

Friedman, D. & Johnson, T. (1988). A mutation in the age-1 gene in Caenorhabditis elegans lengthens life and reduces hermaphrodite fertility. *Genetics*, **118**, 75–86.

Frith, M., Fu, Y., Yu, L., Chen, J., Hansen, U. & Weng, Z. (2004). Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res*, **32**, 1372–81.

Fu, D., Libson, A., Miercke, L., Weitzman, C., Nollert, P., Krucinski, J. & Stroud, R. (2000). Structure of a glycerol-conducting channel and the basis for its selectivity. *Science*, **290**, 481–6.

Furuyama, T., Nakazawa, T., Nakano, I. & Mori, N. (2000). Identification of the differential distribution patterns of mRNAs and consensus binding sequences for mouse DAF-16 homologues. *Biochem J*, **349**, 629–34.

Galibert, M., Carreira, S. & Goding, C. (2001). The Usf-1 transcription factor is a novel target for the stress-responsive p38 kinase and mediates UV-induced tyrosinase expression. *EMBO J*, **20**, 5022–31.

Garlid, K.D., Orosz, D.E., Modriansky, M., Vassanelli, S. & Jezek, P. (1996). On the mechanism of fatty acid-induced proton transport by mitochondrial uncoupling protein. *J Biol Chem*, **271**, 2615–20.

Garlid, K.D., Jaburek, M., Jezek, P. & Varecha, M. (2000). How do uncoupling proteins uncouple? *Biochim Biophys Acta*, **1459**, 383–9.

Garriga, G., Guenther, C. & Horvitz, H. (1993). Migrations of the Caenorhabditis elegans HSNs are regulated by egl-43, a gene encoding two zinc finger proteins. *Genes Dev*, **7**, 2097–109.

Gedik, C., Grant, G., Morrice, P., Wood, S. & Collins, A. (2004). Effects of age and dietary restriction on oxidative DNA damage, antioxidant protection and DNA repair in rats. *Eur J Nutr, Epub ahead of print.*

Gems, D., Sutton, A.J., Sundermeyer, M.L., Albert, P.S., King, K.V., Edgley, M.L., Larsen, P.L. & Riddle, D.L. (1998). Two pleiotropic classes of daf-2 mutation affect larval arrest, adult behavior, reproduction and longevity in Caenorhabditis elegans. *Genetics*, **150**, 129–55.

Gerisch, B., Weitzel, C., Kober-Eisermann, C., Rottiers, V. & Antebi, A. (2001). A hormonal signaling pathway influencing C. elegans metabolism, reproductive development, and life span. *Dev Cell*, **1**, 841–51.

German, M., Wang, J., Chadwick, R. & Rutter, W. (1992). Synergistic activation of the insulin gene by a LIM-homeo domain protein and a basic helix-loop-helix protein: building a functional insulin minienhancer complex. *Genes Dev*, **6**, 2165–76.

Giorgetti, A. & Carloni, P. (2003). Molecular modeling of ion channels: structural predictions. *Curr Opin Chem Biol*, **7**, 150–6.

Girvin, M.E., Rastogi, V.K., Abildgaard, F., Markley, J.L. & Fillingame, R.H. (1998). Solution structure of the transmembrane H+-transporting subunit c of the F1F0 ATP synthase. *Biochemistry*, **37**, 8817–24.

Gius, D., Botero, A., Shah, S. & Curry, H. (1999). Intracellular oxidation/reduction status in the regulation of transcription factors NF-kappaB and AP-1. *Toxicol Lett*, **106**, 93–106.

Golozoubova, V., Hohtola, E., Matthias, A., Jacobsson, A., Cannon, B. & Nedergaard, J. (2001). Only UCP1 can mediate adaptive nonshivering thermogenesis in the cold. *FASEB J*, **15**, 2048–50.

Gong, D.W., He, Y., Karas, M. & Reitman, M. (1997). Uncoupling protein-3 is a mediator of thermogenesis regulated by thyroid hormone, beta3-adrenergic agonists, and leptin. *J Biol Chem*, **272**, 24129–32.

Gong, D.W., Monemdjou, S., Gavrilova, O., Leon, L.R., Marcus-Samuels, B., Chou, C.J., Everett, C., Kozak, L.P., Li, C., Deng, C., Harper, M.E. & Reitman, M.L. (2000). Lack of obesity and normal response to fasting and thyroid hormone in mice lacking uncoupling protein-3. *J Biol Chem*, **275**, 16251–7.

Gonzalez-Barroso, M.M., Fleury, C., Levi-Meyrueis, C., Zaragoza, P., Bouillaud, F. & Rial, E. (1997). Deletion of amino acids 261-269 in the brown fat uncoupling protein converts the carrier into a pore. *Biochemistry*, **36**, 10930–5.

Gonzalez-Barroso, M.M., Fleury, C., Jimenez, M.A., Sanz, J.M., Romero, A., Bouillaud, F. & Rial, E. (1999). Structural and functional study of a conserved region in the uncoupling protein UCP1: the three matrix loops are involved in the control of transport. *J Mol Biol*, **292**, 137–49.

Good, T. & Tatar, M. (2001). Age-specific mortality and reproduction respond to adult dietary restriction in Drosophila melanogaster. *J Insect Physiol*, **47**, 1467–1473.

Gordeliy, V., Labahn, J., Moukhametzianov, R., Efremov, R., Granzin, J., Schlesinger, R., Buldt, G., Savopol, T., Scheidig, A., Klare, J. & Engelhard, M. (2002). Molecular basis of transmembrane signalling by sensory rhodopsin II-transducer complex. *Nature*, **419**, 484–7.

Goueli, B. & Janknecht, R. (2003). Regulation of telomerase reverse transcriptase gene activity by upstream stimulatory factor. *Oncogene*, **22**, 8042–7.

Grav, H., Tronstad, K., Gudbrandsen, O., Berge, K., Fladmark, K., Martinsen, T., Waldum, H., Wergedahl, H. & Berge, R. (2003). Changed energy state and increased mitochondrial beta-oxidation rate in liver of rats associated with lowered proton electrochemical potential and stimulated uncoupling protein 2 (UCP-2) expression: evidence for peroxisome proliferator-activated receptor-alpha independent induction of UCP-2 expression. *J Biol Chem*, **278**, 30525–33.

Gray, T.M. & Matthews, B.W. (1984). Intrahelical hydrogen bonding of serine, threonine and cysteine residues within alpha-helices and its relevance to membrane-bound proteins. *J Mol Biol*, **175**, 75–81.

Greenberg, J.A. & Boozer, C.N. (2000). Metabolic mass, metabolic rate, caloric restriction, and aging in male Fischer 344 rats. *Mech Ageing Dev*, **113**, 37–48.

Grigorieff, N., Ceska, T., Downing, K., Baldwin, J. & Henderson, R. (1996). Electron-crystallographic refinement of the structure of bacteriorhodopsin. *J Mol Biol*, **259**, 393–421.

Guarente, L. & Kenyon, C. (2000). Genetic pathways that regulate ageing in model organisms. *Nature*, **408**, 255–62.

Gueraud, F., Alary, J., Costet, P., Debrauwer, L., Dolo, L., Pineau, T. & Paris, A. (1999). In vivo involvement of cytochrome P450 4A family in the oxidative metabolism of the lipid peroxidation product trans-4-hydroxy-2-nonenal, using PPARalpha-deficient mice. *J Lipid Res*, **40**, 152–9.

GuhaThakurta, D., Palomar, L., Stormo, G., Tedesco, P., Johnson, T., Walker, D., Lithgow, G., Kim, S. & Link, C. (2002). Identification of a novel cis-regulatory element involved in the heat shock response in Caenorhabditis elegans using microarray gene expression and computational methods. *Genome Res*, **12**, 701–12.

Guo, H., Cai, C. & Kuo, P. (2002). Hepatocyte nuclear factor-4alpha mediates redox sensitivity of inducible nitric-oxide synthase gene transcription. *J Biol Chem*, **277**, 5054–60.

Gustafsson, H., Adamson, L., Hedander, J., Walum, E. & Forsby, A. (2001). Insulin-like growth factor type 1 upregulates uncoupling protein 3. *Biochem Biophys Res Commun*, **287**, 1105–11.

Hagen, T. & Lowell, B.B. (2000). Chimeric proteins between UCP1 and UCP3: the middle third of UCP1 is necessary and sufficient for activation by fatty acids. *Biochem Biophys Res Commun*, **276**, 642–8.

Halliwell, B. & Gutteridge, J.M.C. (1999). Free Radicals in Biology and Medicine, Third Edition. *Oxford University Press*.

Hamilton, W. (1966). The moulding of senescence by natural selection. *J Theor Biol*, **12**, 12–45.

Hardenbol, P., Wang, J. & Van Dyke, M. (1997). Identification of preferred hTBP DNA binding sites by the combinatorial method REPSA. *Nucleic Acids Res*, **25**, 3339–44.

Hari, R., Burde, V. & Arking, R. (1998). Immunological confirmation of elevated levels of CuZn superoxide dismutase protein in an artificially selected long-lived strain of Drosophila melanogaster. *Exp Gerontol*, **33**, 227–37.

Harman, D. & Piette, L.H. (1966). Free radical theory of aging: free radical reactions in serum. *J Gerontol*, **21**, 560–5.

Harper, J.A., Stuart, J.A., Jekabsons, M.B., Roussel, D., Brindle, K.M., Dickinson, K., Jones, R.B. & Brand, M.D. (2002). Artifactual uncoupling by uncoupling protein 3 in yeast mitochondria at the concentrations found in mouse and rat skeletal-muscle mitochondria. *Biochem J*, **361**, 49–56.

Harroun, T., Heller, W., Weiss, T., Yang, L. & Huang, H. (1999). Theoretical analysis of hydrophobic matching and membrane-mediated interactions in lipid bilayers containing gramicidin. *Biophys J*, **76**, 3176–85.

Haun, C., Alexander, J., Stainier, D. & Okkema, P. (1998). Rescue of Caenorhabditis elegans pharyngeal development by a vertebrate heart specification gene. *Proc Natl Acad Sci (U S A)*, **95**, 5072–5.

Hegarty, B., Furler, S., Oakes, N., Kraegen, E. & Cooney, G. (2004). Peroxisome proliferator-activated receptor (PPAR) activation induces tissue-specific effects on fatty acid uptake and metabolism in vivo–a study using the novel PPARalpha/gamma agonist tesaglitazar. *Endocrinology*, **145**, 3158–64.

Henderson, S. & Johnson, T. (2001). daf-16 integrates developmental and environmental inputs to mediate aging in the nematode Caenorhabditis elegans. *Curr Biol*, **11**, 1975–80.

Henriksson, M. & Luscher, B. (1996). Proteins of the Myc network: essential regulators of cell growth and differentiation. *Adv Cancer Res*, **68**, 109–82.

Hertweck, M., Gobel, C. & Baumeister, R. (2004). C. elegans SGK-1 is the critical component in the Akt/PKB kinase complex to control stress response and life span. *Dev Cell*, **6**, 577–88.

Himms-Hagen, J. & Harper, M.E. (2001). Physiological role of UCP3 may be export of fatty acids from mitochondria when fatty acid oxidation predominates: an hypothesis. *Exp Biol Med*, **226**, 78–84.

Hirai, H. (1999). The transcription factor Evi-1. *Int J Biochem Cell Biol*, **31**, 1367–71.

Hirai, H., Izutsu, K., Kurokawa, M. & Mitani, K. (2001). Oncogenic mechanisms of Evi-1 protein. *Cancer Chemother Pharmacol*, **48 Suppl 1**, S35–40.

Holzenberger, M., Dupont, J., Ducos, B., Leneuve, P., Geloen, A., Even, P., Cervera, P. & Le Bouc, Y. (2003). IGF-1 receptor regulates lifespan and resistance to oxidative stress in mice. *Nature*, **421**, 182–7.

Honda, Y. & Honda, S. (1999). The daf-2 gene network for longevity regulates oxidative stress resistance and Mn-superoxide dismutase gene expression in Caenorhabditis elegans. *FASEB J*, **13**, 1385–93.

Hope, I., Mounsey, A., Bauer, P. & Aslam, S. (2003). The forkhead gene family of Caenorhabditis elegans. *Gene*, **304**, 43–55.

Hosack, D., Dennis, J.r., Sherman, B., Lane, H. & Lempicki, R. (2003). Identifying biological themes within lists of genes with EASE. *Genome Biol*, **4**, R70.

Hsin, H. & Kenyon, C. (1999). Signals from the reproductive system regulate the lifespan of C. elegans. *Nature*, **399**, 362–6.

Hsu, A., Murphy, C. & Kenyon, C. (2003). Regulation of aging and age-related disease by DAF-16 and heat-shock factor. *Science*, **300**, 1142–5.

Hu, J., Guan, X. & Sanchez, E. (1996). Enhancement of glucocorticoid receptor-mediated gene expression by cellular stress: evidence for the involvement of a heat shock-initiated factor or process during recovery from stress. *Cell Stress Chaperones*, **1**, 197–205.

Huang, S.G., Odoy, S. & Klingenberg, M. (2001). Chimers of two fused ADP/ATP carrier monomers indicate a single channel for ADP/ATP transport. *Arch Biochem Biophys*, **394**, 67–75.

Huang, Y., Lemieux, M., Song, J., Auer, M. & Wang, D. (2003). Structure and mechanism of the glycerol-3-phosphate transporter from Escherichia coli. *Science*, **301**, 616–20.

Hughes, J., Estep, P., Tavazoie, S. & Church, G. (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol*, **296**, 1205–14.

Hunte, C., Koepke, J., Lange, C., Rossmanith, T. & Michel, H. (2000). Structure at 2.3 A resolution of the cytochrome bc(1) complex from the yeast Saccharomyces cerevisiae co-crystallized with an antibody Fv fragment. *Structure Fold Des*, **8**, 669–84.

Ichikawa, M., Asai, T., Chiba, S., Kurokawa, M. & Ogawa, S. (2004). Runx1/AML-1 ranks as a master regulator of adult hematopoiesis. *Cell Cycle*, **3**, 722–4.

Ikeda, K. & Kawakami, K. (1995). DNA binding through distinct domains of zinc-finger-homeodomain protein AREB6 has different effects on gene transcription. *Eur J Biochem*, **233**, 73–82.

Iverson, T.M., Luna-Chavez, C., Cecchini, G. & Rees, D.C. (1999). Structure of the Escherichia coli fumarate reductase respiratory complex. *Science*, **284**, 1961–6.

Iwata, S., Ostermeier, C., Ludwig, B. & Michel, H. (1995). Structure at 2.8 A resolution of cytochrome c oxidase from Paracoccus denitrificans. *Nature*, **376**, 660–9.

Iwata, S., Lee, J.W., Okada, K., Lee, J.K., Iwata, M., Rasmussen, B., Link, T.A., Ramaswamy, S. & Jap, B.K. (1998). Complete structure of the 11-subunit bovine mitochondrial cytochrome bc1 complex. *Science*, **281**, 64–71.

Jaburek, M., Varecha, M., Gimeno, R.E., Dembski, M., Jezek, P., Zhang, M., Burn, P., Tartaglia, L.A. & Garlid, K.D. (1999). Transport function and regulation of mitochondrial uncoupling proteins 2 and 3. *J Biol Chem*, **274**, 26003–7.

Jaiswal, A., Haaparanta, T., Luc, P., Schembri, J. & Adesnik, M. (1990). Glucocorticoid regulation of a phenobarbital-inducible cytochrome P-450 gene: the presence of a functional glucocorticoid response element in the 5'-flanking region of the CYP2B2 gene. *Nucleic Acids Res*, **18**, 4237–42.

Jarmuszkiewicz, W. (2001). Uncoupling proteins in mitochondria of plants and some microorganisms. *Acta Biochim Pol*, **48**, 145–55.

Javadpour, M.M., Eilers, M., Groesbeek, M. & Smith, S.O. (1999). Helix packing in polytopic membrane proteins: role of glycine in transmembrane helix association. *Biophys J*, **77**, 1609–18.

Jayasinghe, S., Hristova, K. & White, S.H. (2001). Energetics, stability, and prediction of transmembrane helices. *J Mol Biol*, **312**, 927–34.

Jezek, P., Modriansky, M. & Garlid, K.D. (1997a). Inactive fatty acids are unable to flip-flop across the lipid bilayer. *FEBS Lett*, **408**, 161–5.

Jezek, P., Modriansky, M. & Garlid, K.D. (1997b). A structure-activity study of fatty acid interaction with mitochondrial uncoupling protein. *FEBS Lett*, **408**, 166–70.

Jezek, P., Lillo, P. & Polecha, J. (1998). Tryptophan fluorescence of mitochondrial uncoupling protein. *Gen Physiol Biophys*, **17**, 157–78.

Jiang, S. & Vakser, I.A. (2000). Side chains in transmembrane helices are shorter at helix-helix interfaces. *Proteins*, **40**, 429–35.

Jiang, Y., Lee, A., Chen, J., Cadene, M., Chait, B. & MacKinnon, R. (2002). Crystal structure and mechanism of a calcium-gated potassium channel. *Nature*, **417**, 515–22.

Jiang, Y., Lee, A., Chen, J., Ruta, V., Cadene, M., Chait, B. & MacKinnon, R. (2003). X-ray structure of a voltage-dependent K+ channel. *Nature*, **423**, 33–41.

Jones, D.T., Taylor, W.R. & Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*, **8**, 275–82.

Jones, D.T., Taylor, W.R. & Thornton, J.M. (1994). A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, **33**, 3038–49.

Jones, P.M. & George, A.M. (2000). Symmetry and structure in P-glycoprotein and ABC transporters what goes around comes around. *Eur J Biochem*, **267**, 5298–305.

Jones, T., Li, D., Wolf, I., Wadekar, S., Periyasamy, S. & Sanchez, E. (2004). Enhancement of glucocorticoid receptor-mediated gene expression by constitutively active heat shock factor 1. *Mol Endocrinol*, **18**, 509–20.

Jordan, P., Fromme, P., Witt, H.T., Klukas, O., Saenger, W. & Krauss, N. (2001). Three-dimensional structure of cyanobacterial photosystem I at 2.5 A resolution. *Nature*, **411**, 909–17.

Jormakka, M., Tornroth, S., Byrne, B. & Iwata, S. (2002). Molecular basis of proton motive force generation: structure of formate dehydrogenase-N. *Science*, **295**, 1863–8.

Jover, R., Bort, R., Gomez-Lechon, M. & Castell, J. (2001). Cytochrome P450 regulation by hepatocyte nuclear factor 4 in human hepatocytes: a study using adenovirus-mediated antisense targeting. *Hepatology*, **33**, 668–75.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–637.

Kairys, V., Gilson, M. & Luy, B. (2004). Structural model for an AxxxG-mediated dimer of surfactant-associated protein C. *Eur J Biochem*, **271**, 2086–92.

Kamiya, N. & Shen, J. (2003). Crystal structure of oxygen-evolving photosystem II from Thermosynechococcus vulcanus at 3.7-A resolution. *Proc Natl Acad Sci U S A*, **100**, 98–103.

Kamp, F., Zakim, D., Zhang, F., Noy, N. & Hamilton, J.A. (1995). Fatty acid flip-flop in phospholipid bilayers is extremely fast. *Biochemistry*, **34**, 11928–37.

Kasprzyk, A., Keefe, D., Smedley, D., London, D., Spooner, W., Melsopp, C., Hammond, M., Rocca-Serra, P., Cox, T. & Birney, E. (2004). EnsMart: a generic system for fast and flexible access to biological data. *Genome Res*, **14**, 160–9.

Katona, G., Andreasson, U., Landau, E., Andreasson, L. & Neutze, R. (2003). Lipidic cubic phase crystal structure of the photosynthetic reaction centre from Rhodobacter sphaeroides at 2.35A resolution. *J Mol Biol*, **331**, 681–92.

Keeton, A., Bortoff, K., Bennett, W., Franklin, J., Venable, D. & Messina, J. (2003). Insulin-regulated expression of Egr-1 and Krox20: dependence on ERK1/2 and interaction with p38 and PI3-kinase pathways. *Endocrinology*, **144**, 5402–10.

Kel, A., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. & Wingender, E. (2003). MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res*, **31**, 3576–9.

Kelly, L.J., Vicario, P.P., Thompson, G.M., Candelore, M.R., Doebber, T.W., Ventre, J., Wu, M.S., Meurer, R., Forrest, M.J., Conner, M.W., Cascieri, M.A. & Moller, D.E. (1998). Peroxisome proliferator-activated receptors gamma and alpha mediate in vivo regulation of uncoupling protein (UCP-1, UCP-2, UCP-3) gene expression. *Endocrinology*, **139**, 4920–7.

Kenyon, C., Chang, J., Gensch, E., Rudner, A. & Tabtiang, R. (1993). A C. elegans mutant that lives twice as long as wild type. *Nature*, **366**, 461–4.

Kerner, J., Turkaly, P.J., Minkler, P.E. & Hoppel, C.L. (2001). Aging skeletal muscle mitochondria in the rat: decreased uncoupling protein-3 content. *Am J Physiol Endocrinol Metab*, **281**, E1054–62.

Kiermaier, A., Gawn, J., Desbarats, L., Saffrich, R., Ansorge, W., Farrell, P., Eilers, M. & Packham, G. (1999). DNA binding of USF is required for specific E-box dependent gene activation in vivo. *Oncogene*, **18**, 7200–11.

Killian, J.A. (1998). Hydrophobic mismatch between proteins and lipids in membranes. *Biochim Biophys Acta*, **1376**, 401–15.

Killian, J.A. & von Heijne, G. (2000). How proteins adapt to a membrane-water interface. *Trends Biochem Sci*, **25**, 429–34.

Kim, S., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J., Eizinger, A., Wylie, B. & Davidson, G. (2001). A gene expression map for Caenorhabditis elegans. *Science*, **293**, 2087–92.

Kim-Han, J.S., Reichert, S.A., Quick, K.L. & Dugan, L.L. (2001). BMCP1: a mitochondrial uncoupling protein in neurons which regulates mitochondrial function and oxidant production. *J Neurochem*, **79**, 658–68.

Kimura, K.D., Tissenbaum, H.A., Liu, Y. & Ruvkun, G. (1997a). daf-2, an insulin receptor-like gene that regulates longevity and diapause in Caenorhabditis elegans. *Science*, **277**, 942–6.

Kimura, Y., Vassylyev, D., Miyazawa, A., Kidera, A., Matsushima, M., Mitsuoka, K., Murata, K., Hirai, T. & Fujiyoshi, Y. (1997b). Surface of bacteriorhodopsin revealed by high-resolution electron crystallography. *Nature*, **389**, 206–11.

Klass, M. (1983). A method for the isolation of longevity mutants in the nematode Caenorhabditis elegans and initial results. *Mech Ageing Dev*, **22**, 279–86.

Klaus, S., Casteilla, L., Bouillaud, F. & Ricquier, D. (1991). The uncoupling protein UCP: a membraneous mitochondrial ion carrier exclusively expressed in brown adipose tissue. *Int J Biochem*, **23**, 791–801.

Klein, C., Garcia-Rizo, C., Bisle, B., Scheffer, B., Zischka, H., Pfeiffer, F., Siedler, F. & Oesterhelt, D. (2004). The membrane proteome of Halobacterium salinarum. *Proteomics*, **5**, 180–197.

Klingenberg, M. (1990). Mechanism and evolution of the uncoupling protein of brown adipose tissue. *Trends Biochem Sci*, **15**, 108–12.

Klingenberg, M. & Appel, M. (1989). The uncoupling protein dimer can form a disulfide cross-link between the mobile C-terminal SH groups. *Eur J Biochem*, **180**, 123–31.

Klingenberg, M. & Echtay, K.S. (2001). Uncoupling proteins: the issues from a biochemist point of view. *Biochim Biophys Acta*, **1504**, 128–43.

Knight, C., Kassen, R., Hebestreit, H. & Rainey, P. (2004). Global analysis of predicted proteomes: functional adaptation of physical properties. *Proc Natl Acad Sci U S A*, **101**, 8390–5.

Koepke, J., Hu, X., Muenke, C., Schulten, K. & Michel, H. (1996). The crystal structure of the light-harvesting complex II (B800-850) from Rhodospirillum molischianum. *Structure*, **4**, 581–97.

Kolbe, M., Besir, H., Essen, L.O. & Oesterhelt, D. (2000). Structure of the light-driven chloride pump halorhodopsin at 1.8 A resolution. *Science*, **288**, 1390–6.

Koopman, P., Munsterberg, A., Capel, B., Vivian, N. & Lovell-Badge, R. (1990). Expression of a candidate sex-determining gene during mouse testis differentiation. *Nature*, **348**, 450–2.

Kotaria, R., Mayor, J.A., Walters, D.E. & Kaplan, R.S. (1999). Oligomeric state of wild-type and cysteine-less yeast mitochondrial citrate transport proteins. *J Bioenerg Biomembr*, **31**, 543–9.

Kozak, L.P. (2000). Uncoupling diet and diabetes. *Nat Med*, **6**, 1092–3.

Kozak, L.P., Britton, J.H., Kozak, U.C. & Wells, J.M. (1988). The mitochondrial uncoupling protein gene. Correlation of exon structure to transmembrane domains. *J Biol Chem*, **263**, 12274–7.

Kuo, A., Gulbis, J., Antcliff, J., Rahman, T., Lowe, E., Zimmer, J., Cuthbertson, J., Ashcroft, F., Ezaki, T. & Doyle, D. (2003). Crystal structure of the potassium channel KirBac1.1 in the closed state. *Science*, **300**, 1922–6.

Kurisu, G., Zhang, H., Smith, J. & Cramer, W. (2003). Structure of the cytochrome b6f complex of oxygenic photosynthesis: tuning the cavity. *Science*, **302**, 1009–14.

Kurokawa, M., Mitani, K., Irie, K., Matsuyama, T., Takahashi, T., Chiba, S., Yazaki, Y., Matsumoto, K. & Hirai, H. (1998). The oncoprotein Evi-1 represses TGF-beta signalling by inhibiting Smad3. *Nature*, **394**, 92–6.

Kyogoku, Y., Fujiyoshi, Y., Shimada, I., Nakamura, H., Tsukihara, T., Akutsu, H., Odahara, T., Okada, T. & Nomura, N. (2003). Structural genomics of membrane proteins. *Acc Chem Res*, **36**, 199–206.

Kyte, J. & Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, **157**, 105–32.

Lancaster, C., Kroger, A., Auer, M. & Michel, H. (1999). Structure of fumarate reductase from Wolinella succinogenes at 2.2 A resolution. *Nature*, **402**, 377–85.

Landfield, P. (1994). Nathan Shock Memorial Lecture 1990. The role of glucocorticoids in brain aging and Alzheimer's disease: an integrative physiological hypothesis. *Exp Gerontol*, **29**, 3–11.

Langosch, D. & Heringa, J. (1998). Interaction of transmembrane helices by a knobs-into-holes packing characteristic of soluble coiled coils. *Proteins*, **31**, 150–9.

Larkin, S., Mull, E., Miao, W., Pittner, R., Albrandt, K., Moore, C., Young, A., Denaro, M. & Beaumont, K. (1997). Regulation of the third member of the uncoupling protein family, UCP3, by cold and thyroid hormone. *Biochem Biophys Res Commun*, **240**, 222–7.

Lasky, L. (1994). Sialomucin ligands for selectins: a new family of cell adhesion molecules. *Princess Takamatsu Symp*, **24**, 81–90.

Lasky, L. (1995). Selectin-carbohydrate interactions and the initiation of the inflammatory response. *Annu Rev Biochem*, **64**, 113–39.

LaVoie, H. (2003). The role of GATA in mammalian reproduction. *Exp Biol Med*, **228**, 1282–90.

Laws, T., Harding, S., Smith, M., Atkins, T. & Titball, R. (2004). Age influences resistance of Caenorhabditis elegans to killing by pathogenic bacteria. *FEMS Microbiol Lett*, **234**, 281–7.

Ledesma, A., de Lacoba, M.G., Arechaga, I. & Rial, E. (2002). Modeling the transmembrane arrangement of the uncoupling protein UCP1 and topological considerations of the nucleotide-binding site. *J Bioenerg Biomembr*, **34**, 473–86.

Lee, B. & Richards, F.M. (1971). The interpretation of protein structures: estimation of static accessibility. *J Mol Biol*, **55**, 379–400.

Lee, S., Kennedy, S., Tolonen, A. & Ruvkun, G. (2003). DAF-16 target genes that control C. elegans life-span and metabolism. *Science*, **300**, 644–7.

Lee, S., Shah, S., Yu, C., Wigley, W., Li, H., Lim, M., Pedersen, K., Han, W., Thomas, P., Lundkvist, J., Hao, Y. & Yu, G. (2004). A conserved GXXXG motif in APH-1 is critical for assembly and activity of the gamma-secretase complex. *J Biol Chem*, **279**, 4144–52.

Li, C. & Tucker, P. (1993). DNA-binding properties and secondary structural model of the hepatocyte nuclear factor 3/fork head domain. *Proc Natl Acad Sci (U S A)*, **90**, 11583–7.

Li, D., Periyasamy, S., Jones, T. & Sanchez, E. (2000). Heat and chemical shock potentiation of glucocorticoid receptor transactivation requires heat shock factor (HSF) activity. Modulation of HSF by vanadate and wortmannin. *J Biol Chem*, **275**, 26058–65.

Li, G., Currie, R. & Ali, I. (2004a). Insulin potentiates expression of myocardial heat shock protein 70. *Eur J Cardiothorac Surg*, **26**, 281–8.

Li, Q., Hu, N., Daggett, M., Chu, W., Bittel, D., Johnson, J. & Andrews, G. (1998). Participation of upstream stimulator factor (USF) in cadmium-induction of the mouse metallothionein-I gene. *Nucleic Acids Res*, **26**, 5182–9.

Li, R., Gorelik, R., Nanda, V., Law, P., Lear, J., DeGrado, W. & Bennett, J. (2004b). Dimerization of the transmembrane domain of integrin alphaIIb subunit in cell membranes. *J Biol Chem*, **279**, 26666–26673.

Li, S.C. & Deber, C.M. (1992). Glycine and beta-branched residues support and modulate peptide helicity in membrane environments. *FEBS Lett*, **311**, 217–20.

Libina, N., Berman, J. & Kenyon, C. (2003). Tissue-specific activities of C. elegans DAF-16 in the regulation of lifespan. *Cell*, **115**, 489–502.

Lin, C.S., Hackenberg, H. & Klingenberg, E.M. (1980). The uncoupling protein from brown adipose tissue mitochondria is a dimer. A hydrodynamic study. *FEBS Lett*, **113**, 304–6.

Lin, K., Hsin, H., Libina, N. & Kenyon, C. (2001). Regulation of the Caenorhabditis elegans longevity protein DAF-16 by insulin/IGF-1 and germline signaling. *Nat Genet*, **28**, 139–45.

Liu, F., Lewis, R., Hodges, R. & McElhaney, R. (2004a). Effect of variations in the structure of a polyleucine-based alpha-helical transmembrane peptide on its interaction with phosphatidylethanolamine Bilayers. *Biophys J*, **87**, 2470–82.

Liu, F., Lewis, R., Hodges, R. & McElhaney, R. (2004b). Effect of variations in the structure of a polyleucine-based alpha-helical transmembrane peptide on its interaction with phosphatidylglycerol bilayers. *Biochemistry*, **43**, 3679–87.

Liu, Q., Bai, C., Chen, F., Wang, R., MacDonald, T., Gu, M., Zhang, Q., Morsy, M.A. & Caskey, C.T. (1998). Uncoupling protein-3: a muscle-specific gene upregulated by leptin in ob/ob mice. *Gene*, **207**, 1–7.

Liu, W., Eilers, M., Patel, A. & Smith, S. (2004c). Helix packing moments reveal diversity and conservation in membrane protein structure. *J Mol Biol*, **337**, 713–29.

Liu, Y., Engelman, D. & Gerstein, M. (2002). Genomic analysis of membrane protein families: abundance and conserved motifs. *Genome Biol*, **3**, research0054.

Locher, K.P., Lee, A.T. & Rees, D.C. (2002). The E. coli BtuCD structure: a framework for ABC transporter architecture and mechanism. *Science*, **296**, 1091–8.

Luckinbil, L.S., Arking, M.J., Clare, M.J., Cirocco, W.C. & Buck, S.A. (1984). Selection for delayed senescence in Drosophila melanogaster. *Evolution*, **38**, 966–1003.

Ludewig, A., Kober-Eisermann, C., Weitzel, C., Bethke, A., Neubert, K., Gerisch, B., Hutter, H. & Antebi, A. (2004). A novel nuclear receptor/coregulator complex controls C. elegans lipid metabolism, larval development, and aging. *Genes Dev*, **18**, 2120–33.

Luecke, H., Richter, H. & Lanyi, J. (1998). Proton transfer pathways in bacteriorhodopsin at 2.3 angstrom resolution. *Science*, **280**, 1934–7.

Luecke, H., Schobert, B., Richter, H., Cartailler, J. & Lanyi, J. (1999a). Structural changes in bacteriorhodopsin during ion transport at 2 angstrom resolution. *Science*, **286**, 255–61.

Luecke, H., Schobert, B., Richter, H., Cartailler, J. & Lanyi, J. (1999b). Structure of bacteriorhodopsin at 1.55 A resolution. *J Mol Biol*, **291**, 899–911.

Luecke, H., Schobert, B., Lanyi, J., Spudich, E. & Spudich, J. (2001). Crystal structure of sensory rhodopsin II at 2.4 angstroms: insights into color tuning and transducer interaction. *Science*, **293**, 1499–503.

Ma, L., Mao, S., Taylor, K., Kanjanabuch, T., Guan, Y., Zhang, Y., Brown, N., Swift, L., McGuinness, O., Wasserman, D., Vaughan, D. & Fogo, A. (2004). Prevention of obesity and insulin resistance in mice lacking plasminogen activator inhibitor 1. *Diabetes*, **53**, 336–46.

Maglich, J., Sluder, A., Guan, X., Shi, Y., McKee, D., Carrick, K., Kamdar, K., Willson, T. & Moore, J. (2001). Comparison of complete nuclear receptor sets from the human, Caenorhabditis elegans and Drosophila genomes. *Genome Biol*, **2**, 29–31.

Malashkevich, V.N., Chan, D.C., Chutkowski, C.T. & Kim, P.S. (1998). Crystal structure of the simian immunodeficiency virus (SIV) gp41 core: conserved helical interactions underlie the broad inhibitory activity of gp41 peptides. *Proc Natl Acad Sci (U S A)*, **95**, 9134–9.

Mao, W., Yu, X.X., Zhong, A., Li, W., Brush, J., Sherwood, S.W., Adams, S.H. & Pan, G. (1999). UCP4, a novel brain-specific mitochondrial protein that reduces membrane potential in mammalian cells. *FEBS Lett*, **443**, 326–30.

Margalit, Y., Yarus, S., Shapira, E., Gruenbaum, Y. & Fainsod, A. (1993). Isolation and characterization of target sequences of the chicken CdxA homeobox gene. *Nucleic Acids Res*, **21**, 4915–22.

Marqusee, S. & Baldwin, R.L. (1987). Helix stabilization by Glu-...Lys+ salt bridges in short peptides of de novo design. *Proc Natl Acad Sci (U S A)*, **84**, 8898–902.

Masoro, E.J. (1995). Glucocorticoids and aging. *Aging (Milano)*, **7**, 407–13.

Masoro, E.J. & Austad, S.N. (1996). The evolution of the antiaging action of dietary restriction: a hypothesis. *J Gerontol A Biol Sci Med Sci*, **51**, B387–91.

Mayinger, P. & Klingenberg, M. (1992). Labeling of two different regions of the nucleotide binding site of the uncoupling protein from brown adipose tissue mitochondria with two ATP analogs. *Biochemistry*, **31**, 10536–43.

McClain, M., Iwamoto, H., Cao, P., Vinion-Dubiel, A., Li, Y., Szabo, G., Shao, Z. & Cover, T. (2003). Essential role of a GXXXG motif for membrane channel formation by Helicobacter pylori vacuolating toxin. *J Biol Chem*, **278**, 12101–8.

McDermott, G., Prince, S., Freer, A., Hawthornthwaite-Lawless, A., Papiz, M., Cogdell, R. & Isaacs, N. (1995). Crystal structure of an integral membrane light-harvesting complex from photosynthetic bacteria. *Nature*, **374**, 517–521.

McDonald, I.K. & Thornton, J.M. (1994). Satisfying hydrogen bonding potential in proteins. *J Mol Biol*, **238**, 777–93.

McElwee, J., Schuster, E., Blanc, E., Thomas, J. & Gems, D. (2004). Shared transcriptional signature in Caenorhabditis elegans Dauer larvae and long-lived daf-2 mutants implicates detoxification system in longevity assurance. *J Biol Chem*, **279**, 44533–43.

Medvedev, A.V., Snedden, S.K., Raimbault, S., Ricquier, D. & Collins, S. (2001). Transcriptional regulation of the mouse uncoupling protein-2 gene. Double E-box motif is required for peroxisome proliferator-activated receptor-gamma-dependent activation. *J Biol Chem*, **276**, 10817–23.

Mendrola, J., Berger, M., King, M. & Lemmon, M. (2002). The single transmembrane domains of ErbB receptors self-associate in cell membranes. *J Biol Chem*, **277**, 4704–12.

Milks, L., Kumar, N., Houghten, R., Unwin, N. & Gilula, N. (1988). Topology of the 32-kd liver gap junction protein determined by site-directed antibody localizations. *EMBO J*, **7**, 2967–75.

Miroux, B., Casteilla, L., Klaus, S., Raimbault, S., Grandin, S., Clement, J.M., Ricquier, D. & Bouillaud, F. (1992). Antibodies selected from whole antiserum by fusion proteins as tools for the study of the topology of mitochondrial membrane proteins. Evidence that the N-terminal extremity of the sixth alpha-helix of the uncoupling protein is facing the matrix. *J Biol Chem*, **267**, 13603–9.

Miroux, B., Frossard, V., Raimbault, S., Ricquier, D. & Bouillaud, F. (1993). The topology of the brown adipose tissue mitochondrial uncoupling protein determined with antibodies against its antigenic sites revealed by a library of fusion proteins. *EMBO J*, **12**, 3739–45.

Mishra, V.K., Palgunachari, M.N., Segrest, J.P. & Anantharamaiah, G.M. (1994). Interactions of synthetic peptide analogs of the class A amphipathic helix with lipids. Evidence for the snorkel hypothesis. *J Biol Chem*, **269**, 7185–91.

Miyazawa, A., Fujiyoshi, Y. & Unwin, N. (2003). Structure and gating mechanism of the acetylcholine receptor pore. *Nature*, **423**, 949–55.

Mizuno, T., Miura-Suzuki, T., Yamashita, H. & Mori, N. (2000). Distinct regulation of brain mitochondrial carrier protein-1 and uncoupling protein-2 genes in the rat brain during cold exposure and aging. *Biochem Biophys Res Commun*, **278**, 691–7.

Modriansky, M., Murdza-Inglis, D.L., Patel, H.V., Freeman, K.B. & Garlid, K.D. (1997). Identification by site-directed mutagenesis of three arginines in uncoupling protein that are essential for nucleotide binding and inhibition. *J Biol Chem*, **272**, 24759–62.

Moller, D. & Berger, J. (2003). Role of PPARs in the regulation of obesity-related insulin sensitivity and inflammation. *Int J Obes Relat Metab Disord*, **27 Suppl 3**, S17–21.

Moller, S., Croning, M.D. & Apweiler, R. (2001). Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, **17**, 646–53.

Moore, D., Marks, A., Buckley, D., Kapler, G., Payvar, F. & Goodman, H. (1985). The first intron of the human growth hormone gene contains a binding site for glucocorticoid receptor. *Proc Natl Acad Sci (U S A)*, **82**, 699–702.

Morishita, K., Parganas, E., Parham, D., Matsugi, T. & Ihle, J. (1990). The Evi-1 zinc finger myeloid transforming gene is normally expressed in the kidney and in developing oocytes. *Oncogene*, **5**, 1419–23.

Morris, J., Tissenbaum, H. & Ruvkun, G. (1996). A phosphatidylinositol-3-OH kinase family member regulating longevity and diapause in Caenorhabditis elegans. *Nature*, **382**, 536–9.

Muller, V., Basset, G., Nelson, D.R. & Klingenberg, M. (1996). Probing the role of positive residues in the ADP/ATP carrier from yeast. The effect of six arginine mutations of oxidative phosphorylation and AAC expression. *Biochemistry*, **35**, 16132–43.

Murakami, S. & Johnson, T. (1996). A genetic pathway conferring life extension and resistance to UV stress in Caenorhabditis elegans. *Genetics*, **143**, 1207–18.

Murakami, S., Nakashima, R., Yamashita, E. & Yamaguchi, A. (2002). Crystal structure of bacterial multidrug efflux transporter AcrB. *Nature*, **419**, 587–93.

Murata, K., Mitsuoka, K., Hirai, T., Walz, T., Agre, P., Heymann, J., Engel, A. & Fujiyoshi, Y. (2000). Structural determinants of water permeation through aquaporin-1. *Nature*, **407**, 599–605.

Murphy, C., McCarroll, S., Bargmann, C., Fraser, A., Kamath, R., Ahringer, J., Li, H. & Kenyon, C. (2003). Genes that act downstream of DAF-16 to influence the lifespan of Caenorhabditis elegans. *Nature*, **424**, 277–83.

Nedellec, P., Edling, Y., Perret, E., Fardeau, M. & Vicart, P. (2002). Glucocorticoid treatment induces expression of small heat shock proteins in human satellite cell populations:

consequences for a desmin-related myopathy involving the R120G alpha B-crystallin mutation. *Neuromuscul Disord*, **12**, 457–65.

Nelson, D.R. & Douglas, M.G. (1993). Function-based mapping of the yeast mitochondrial ADP/ATP translocator by selection for second site revertants. *J Mol Biol*, **230**, 1171–82.

Nibbelink, M., Moulin, K., Arnaud, E., Duval, C., Penicaud, L. & Casteilla, L. (2001). Brown fat UCP1 is specifically expressed in uterine longitudinal smooth muscle cells. *J Biol Chem*, **276**, 47291–5.

Nikiforovich, G.V., Galaktionov, S., Balodis, J. & Marshall, G.R. (2001). Novel approach to computer modeling of seven-helical transmembrane proteins: current progress in the test case of bacteriorhodopsin. *Acta Biochim Pol*, **48**, 53–64.

Nilsson, I., Saaf, A., Whitley, P., Gafvelin, G., Waller, C. & von Heijne, G. (1998). Proline-induced disruption of a transmembrane alpha-helix in its natural environment. *J Mol Biol*, **284**, 1165–75.

Nishida, M. & MacKinnon, R. (2002). Structural basis of inward rectification: cytoplasmic pore of the G protein-gated inward rectifier GIRK1 at 1.8 A resolution. *Cell*, **111**, 957–65.

Nishikawa, T., Edelstein, D., Du, X.L., Yamagishi, S., Matsumura, T., Kaneda, Y., Yorek, M.A., Beebe, D., Oates, P.J., Hammes, H.P., Giardino, I. & Brownlee, M. (2000). Normalizing mitochondrial superoxide production blocks three pathways of hyperglycaemic damage. *Nature*, **404**, 787–90.

Nogi, T., Fathir, I., Kobayashi, M., Nozawa, T. & Miki, K. (2000). Crystal structures of photosynthetic reaction center and high-potential iron-sulfur protein from Thermochromatium tepidum: thermostability and electron transfer. *Proc Natl Acad Sci U S A*, **97**, 13561–6.

Noselli, S., Payre, F. & Vincent, A. (1992). Zinc fingers and other domains cooperate in binding of Drosophila sry beta and delta proteins at specific chromosomal sites. *Mol Cell Biol*, **12**, 724–33.

Ogg, S., Paradis, S., Gottlieb, S., Patterson, G., Lee, L., Tissenbaum, H. & Ruvkun, G. (1997). The Fork head transcription factor DAF-16 transduces insulin-like metabolic and longevity signals in C. elegans. *Nature*, **389**, 994–9.

Oiki, S., Madison, V. & Montal, M. (1990). Bundles of amphipathic transmembrane alpha-helices as a structural motif for ion-conducting channel proteins: studies on sodium channels and acetylcholine receptors. *Proteins*, **8**, 226–36.

Okada, T., Fujiyoshi, Y., Silow, M., Navarro, J., Landau, E. & Shichida, Y. (2002). Functional role of internal water molecules in rhodopsin revealed by X-ray crystallography. *Proc Natl Acad Sci U S A*, **99**, 5982–7.

O'Keeffe, A., East, J. & Lee, A. (2000). Selectivity in lipid binding to the bacterial outer membrane protein OmpF. *Biophys J*, **79**, 2066–74.

Okkema, P. & Fire, A. (1994). The Caenorhabditis elegans NK-2 class homeoprotein CEH-22 is involved in combinatorial activation of gene expression in pharyngeal muscle. *Development*, **120**, 2175–86.

Omiecinski, C., Remmel, R. & Hosagrahara, V. (1999). Concise review of the cytochrome P450s and their roles in toxicology. *Toxicol Sci*, **48**, 151–6.

Ookuma, S., Fukuda, M. & Nishida, E. (2003). Identification of a DAF-16 transcriptional target gene, scl-1, that regulates longevity and stress resistance in Caenorhabditis elegans. *Curr Biol*, **13**, 427–31.

Opella, S.J., Marassi, F.M., Gesell, J.J., Valente, A.P., Kim, Y., Oblatt-Montal, M. & Montal, M. (1999). Structures of the M2 channel-lining segments from nicotinic acetylcholine and NMDA receptors by NMR spectroscopy. *Nat Struct Biol*, **6**, 374–9.

O'Riordan, V. & Burnell, A. (1990). Interemediary metabolism in the dauer larva of the nematode Caenorhabditis elegans- II. The glyoxylate cycle and fatty-acid oxidation. *Comp Biochem Phys B*, **95**, 125–130.

Orr, W. & Sohal, R. (1992). The effects of catalase gene overexpression on life span and resistance to oxidative stress in transgenic Drosophila melanogaster. *Arch Biochem Biophys*, **297**, 35–41.

Orr, W. & Sohal, R. (1993). Effects of Cu-Zn superoxide dismutase overexpression of life span and resistance to oxidative stress in transgenic Drosophila melanogaster. *Arch Biochem Biophys*, **301**, 34–40.

Orr, W. & Sohal, R. (1994). Extension of life-span by overexpression of superoxide dismutase and catalase in Drosophila melanogaster. *Science*, **263**, 1128–30.

Orr, W. & Sohal, R. (2003). Does overexpression of Cu,Zn-SOD extend life span in Drosophila melanogaster? *Exp Gerontol*, **38**, 227–30.

Overton, M., Chinault, S. & Blumer, K. (2003). Oligomerization, biogenesis, and signaling is promoted by a glycophorin A-like dimerization motif in transmembrane domain 1 of a yeast G protein-coupled receptor. *J Biol Chem*, **278**, 49369–77.

Palczewski, K., Kumasaka, T., Hori, T., Behnke, C.A., Motoshima, H., Fox, B.A., Le Trong, I., Teller, D.C., Okada, T., Stenkamp, R.E., Yamamoto, M. & Miyano, M. (2000). Crystal structure of rhodopsin: A G protein-coupled receptor. *Science*, **289**, 739–45.

Papatsenko, D., Makeev, V., Lifanov, A., Regnier, M., Nazina, A. & Desplan, C. (2002). Extraction of functional binding sites from unique regulatory regions: the Drosophila early developmental enhancers. *Genome Res*, **12**, 470–81.

Papavassiliou, A. (2001). Determination of a transcription-factor-binding site by nuclease protection footprinting onto southwestern blots. *Methods Mol Biol*, **148**, 135–49.

Papiz, M., Prince, S., Howard, T., Cogdell, R. & Isaacs, N. (2003). The structure and thermal motion of the B800-850 LH2 complex from Rps.acidophila at 2.0A resolution and 100K: new structural features and functionally relevant motions. *J Mol Biol*, **326**, 1523–38.

Parkes, T.L., Elia, A.J., Dickinson, D., Hilliker, A.J., Phillips, J.P. & Boulianne, G.L. (1998a). Extension of Drosophila lifespan by overexpression of human SOD1 in motorneurons. *Nat Genet*, **19**, 171–4.

Parkes, T.L., Kirby, K., Phillips, J.P. & Hilliker, A.J. (1998b). Transgenic analysis of the cSOD-null phenotypic syndrome in Drosophila. *Genome*, **41**, 642–51.

Parkinson, D., Bhaskaran, A., Droggiti, A., Dickinson, S., D'Antonio, M., Mirsky, R. & Jessen, K. (2004). Krox-20 inhibits Jun-NH2-terminal kinase/c-Jun to control Schwann cell proliferation and death. *J Cell Biol*, **164**, 385–94.

Partridge, L. & Gems, D. (2002). Mechanisms of ageing: public or private? *Nature Rev Genet*, **3**, 165–175.

Patient, R. & McGhee, J. (2002). The GATA family (vertebrates and invertebrates). *Curr Opin Genet Dev*, **12**, 416–22.

Paulson, K., Darnell, J.r., Rushmore, T. & Pickett, C. (1990). Analysis of the upstream elements of the xenobiotic compound-inducible and positionally regulated glutathione S-transferase Ya gene. *Mol Cell Biol*, **10**, 1841–52.

Payre, F. & Vincent, A. (1991). Genomic targets of the serendipity beta and delta zinc finger proteins and their respective DNA recognition sites. *EMBO J*, **10**, 2533–41.

Pebay-Peyroula, E., Rummel, G., Rosenbusch, J. & Landau, E. (1997). X-ray structure of bacteriorhodopsin at 2.5 angstroms from microcrystals grown in lipidic cubic phases. *Science*, **277**, 1676–81.

Pebay-Peyroula, E., Dahout-Gonzalez, C., Kahn, R., Trezeguet, V., Lauquin, G. & Brandolin, G. (2003). Structure of mitochondrial ADP/ATP carrier in complex with carboxyatractyloside. *Nature*, **426**, 39–44.

Pecqueur, C., Cassard-Doulcier, A.M., Raimbault, S., Miroux, B., Fleury, C., Gelly, C., Bouillaud, F. & Ricquier, D. (1999). Functional organization of the human uncoupling protein-2 gene, and juxtaposition to the uncoupling protein-3 gene. *Biochem Biophys Res Commun*, **255**, 40–6.

Pecqueur, C., Alves-Guerra, M.C., Gelly, C., Levi-Meyrueis, C., Couplan, E., Collins, S., Ricquier, D., Bouillaud, F. & Miroux, B. (2001). Uncoupling protein 2, in vivo distribution, induction upon oxidative stress, and evidence for translational regulation. *J Biol Chem*, **276**, 8705–12.

Pedersen, S.B., Bruun, J.M., Kristensen, K. & Richelsen, B. (2001). Regulation of UCP1, UCP2, and UCP3 mRNA expression in brown adipose tissue, white adipose tissue, and skeletal muscle in rats by estrogen. *Biochem Biophys Res Commun*, **288**, 191–7.

Pellegrini-Calace, M., Carotti, A. & Jones, D.T. (2003). Folding in lipid membranes (FILM): A novel method for the prediction of small membrane protein 3D structures. *Proteins*, **50**, 537–45.

Perkins, A., Mercer, J., Jenkins, N. & Copeland, N. (1991). Patterns of Evi-1 expression in embryonic and adult tissues suggest that Evi-1 plays an important regulatory role in mouse development. *Development*, **111**, 479–87.

Phillips, J., Campbell, S., Michaud, D., Charbonneau, M. & Hilliker, A. (1989). Null mutation of copper/zinc superoxide dismutase in Drosophila confers hypersensitivity to paraquat and reduced longevity. *Proc Natl Acad Sci (U S A)*, **86**, 2761–5.

Pierce, S.B., Costa, M., Wisotzkey, R., Devadhar, S., Homburger, S.A., Buchman, A.R., Ferguson, K.C., Heller, J., Platt, D.M., Pasquinelli, A.A., Liu, L.X., Doberstein, S.K. & Ruvkun, G. (2001). Regulation of DAF-2 receptor signaling by human insulin and ins-1, a member of the unusually large and diverse C. elegans insulin gene family. *Genes Dev*, **15**, 672–86.

Pilpel, Y., Ben-Tal, N. & Lancet, D. (1999). kPROT: a knowledge-based scale for the propensity of residue orientation in transmembrane segments. Application to membrane protein structure prediction. *J Mol Biol*, **294**, 921–35.

Polgar, O., Robey, R., Morisaki, K., Dean, M., Michejda, C., Sauna, Z., Ambudkar, S., Tarasova, N. & Bates, S. (2004). Mutational analysis of ABCG2: role of the GXXXG motif. *Biochemistry*, **43**, 9448–56.

Prado, C., Pugh-Bernard, A., Elghazi, L., Sosa-Pineda, B. & Sussel, L. (2004). Ghrelin cells replace insulin-producing beta cells in two mouse models of pancreas development. *Proc Natl Acad Sci (U S A)*, **101**, 2924–9.

Prestridge, D. (1996). SIGNAL SCAN 4.0: additional databases and sequence formats. *Comput Appl Biosci*, **12**, 157–60.

Quandt, K., Frech, K., Karas, H., Wingender, E. & Werner, T. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res*, **23**, 4878–84.

Rahman, I., Gilmour, P., Jimenez, L. & MacNee, W. (2002). Oxidative stress and TNF-alpha induce histone acetylation and NF-kappaB/AP-1 activation in alveolar epithelial cells: potential mechanism in gene transcription in lung inflammation. *Mol Cell Biochem*, **234-235**, 239–48.

Rees, D.C. & Eisenberg, D. (2000). Turning a reference inside-out: commentary on an article by Stevens and Arkin entitled: "Are membrane proteins 'inside-out' proteins?" (Proteins 1999;36:135-143). *Proteins*, **38**, 121–2.

Rees, D.C., DeAntonio, L. & Eisenberg, D. (1989). Hydrophobic organization of membrane proteins. *Science*, **245**, 510–3.

Reid, K. & Nelson, C. (2001). Improved methylation protection-based DNA footprinting to reveal structural distortion of DNA upon transcription factor binding. *Biotechniques*, **30**, 20–2.

Ren, G., Reddy, V., Cheng, A., Melnyk, P. & Mitra, A. (2001). Visualization of a water-selective pore by electron crystallography in vitreous ice. *Proc Natl Acad Sci U S A*, **98**, 1398–403.

Rial, E., Gonzalez-Barroso, M., Fleury, C., Iturrizaga, S., Sanchis, D., Jimenez-Jimenez, J., Ricquier, D., Goubern, M. & Bouillaud, F. (1999). Retinoids activate proton transport by the uncoupling proteins UCP1 and UCP2. *EMBO J*, **18**, 5827–33.

Rojo, E.E., Guiard, B., Neupert, W. & Stuart, R.A. (1999). N-terminal tail export from the mitochondrial matrix. Adherence to the prokaryotic "positive-inside" rule of membrane protein topology. *J Biol Chem*, **274**, 19617–22.

Rosen, S. (1993). Cell surface lectins in the immune system. *Semin Immunol*, **5**, 237–47.

Roth, F., Hughes, J., Estep, P. & Church, G. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol*, **16**, 939–45.

Royant, A., Nollert, P., Edman, K., Neutze, R., Landau, E., Pebay-Peyroula, E. & Navarro, J. (2001). X-ray structure of sensory rhodopsin II at 2.1-A resolution. *Proc Natl Acad Sci U S A*, **98**, 10131–6.

Runswick, M.J., Powell, S.J., Nyren, P. & Walker, J.E. (1987). Sequence of the bovine mitochondrial phosphate carrier protein: structural relationship to ADP/ATP translocase and the brown fat mitochondria uncoupling protein. *EMBO J*, **6**, 1367–73.

Sabatino, F., Masoro, E.J., McMahan, C.A. & Kuhn, R.W. (1991). Assessment of the role of the glucocorticoid system in aging processes and in the action of food restriction. *J Gerontol*, **46**, B171–9.

Sanchez, E., Hu, J., Zhong, S., Shen, P., Greene, M. & Housley, P. (1994). Potentiation of glucocorticoid receptor-mediated gene expression by heat and chemical shock. *Mol Endocrinol*, **8**, 408–21.

Sanchis, D., Fleury, C., Chomiki, N., Goubern, M., Huang, Q., Neverova, M., Gregoire, F., Easlick, J., Raimbault, S., Levi-Meyrueis, C., Miroux, B., Collins, S., Seldin, M., Richard, D., Warden, C., Bouillaud, F. & Ricquier, D. (1998). BMCP1, a novel mitochondrial carrier with high expression in the central nervous system of humans and rodents, and respiration uncoupling activity in recombinant yeast. *J Biol Chem*, **273**, 34611–5.

Sander, M., Neubuser, A., Kalamaras, J., Ee, H., Martin, G. & German, M. (1997). Genetic analysis reveals that PAX6 is required for normal transcription of pancreatic hormone genes and islet development. *Genes Dev*, **11**, 1662–73.

Sapolsky, R. (1992). Do glucocorticoid concentrations rise with age in the rat? *Neurobiol Aging*, **13**, 171–4.

Sapolsky, R. (1999). Glucocorticoids, stress, and their adverse neurological effects: relevance to aging. *Exp Gerontol*, **34**, 721–32.

Sapolsky, R., Armanini, M., Packan, D. & Tombaugh, G. (1987). Stress and glucocorticoids in aging. *Endocrinol Metab Clin North Am*, **16**, 965–80.

Savage, D., Egea, P., Robles-Colmenares, Y., O'Connell, .r.d. & Stroud, R. (2003). Architecture and selectivity in aquaporins: 2.5 a X-ray structure of aquaporin Z. *PLoS Biol*, **1**, E72.

Schiffer, M. & Edmundson, A. (1967). Use of helical wheels to represent the structures of proteins and to identify segments with helical potential. *Biophys J*, **7**, 121–35.

Schiffer, M., Ainsworth, C.F., Deng, Y.L., Johnson, G., Pascoe, F.H. & Hanson, D.K. (1995). Proline in a transmembrane helix compensates for cavities in the photosynthetic reaction center. *J Mol Biol*, **252**, 472–82.

Schrauwen, P., Walder, K. & Ravussin, E. (1999). Human uncoupling proteins and obesity. *Obes Res*, **7**, 97–105.

Schrauwen, P., Hesselink, M.K., Blaak, E.E., Borghouts, L.B., Schaart, G., Saris, W.H. & Keizer, H.A. (2001). Uncoupling protein 3 content is decreased in skeletal muscle of patients with type 2 diabetes. *Diabetes*, **50**, 2870–3.

Schroers, A., Burkovski, A., Wohlrab, H. & Kramer, R. (1998). The phosphate carrier from yeast mitochondria. Dimerization is a prerequisite for function. *J Biol Chem*, **273**, 14269–76.

Schubert, W., Klukas, O., Krauss, N., Saenger, W., Fromme, P. & Witt, H. (1997). Photosystem I of Synechococcus elongatus at 4 A resolution: comprehensive structure analysis. *J Mol Biol*, **272**, 741–69.

Schuenemann, T.A., Delgado-Nixon, V.M. & Dalbey, R.E. (1999). Direct evidence that the proton motive force inhibits membrane translocation of positively charged residues within membrane proteins. *J Biol Chem*, **274**, 6855–64.

Schwartz, R., Ting, C. & King, J. (2001). Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of life. *Genome Res*, **11**, 703–9.

Schwarze, S.R., Weindruch, R. & Aiken, J.M. (1998). Oxidative stress and aging reduce COX I RNA and cytochrome oxidase activity in Drosophila. *Free Radic Biol Med*, **25**, 740–7.

Schwede, T., Kopp, J., Guex, N. & Peitsch, M. (2003). SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res*, **31**, 3381–5.

Semenza, G. (2000). Surviving ischemia: adaptive responses mediated by hypoxia-inducible factor 1. *J Clin Invest*, **106**, 809–12.

Semenza, G. (2004). Hydroxylation of HIF-1: oxygen sensing at the molecular level. *Physiology*, **19**, 176–82.

Senes, A., Gerstein, M. & Engelman, D.M. (2000). Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. *J Mol Biol*, **296**, 921–36.

Senes, A., Ubarretxena-Belandia, I. & Engelman, D.M. (2001). The Calpha —H...O hydrogen bond: a determinant of stability and specificity in transmembrane helix interactions. *Proc Natl Acad Sci (U S A)*, **98**, 9056–61.

Seto, N., Hayashi, S. & Tener, G. (1990). Overexpression of Cu-Zn superoxide dismutase in Drosophila does not affect life-span. *Proc Natl Acad Sci (U S A)*, **87**, 4270–4.

Sgro, C. & Partridge, L. (1999). A delayed wave of death from reproduction in Drosophila. *Science*, **286**, 2521–4.

Shah, V. & Smart, V. (1996). Human chromosome Y and SRY. *Cell Biol Int*, **20**, 3–6.

Shaw, P. & Stewart, A. (1994). Identification of protein-DNA contacts with dimethyl sulfate. Methylation protection and methylation interference. *Methods Mol Biol*, **30**, 79–87.

Shen, Q., Cline, G., Shulman, G., Leibowitz, M. & Davies, P. (2004). Effects of rexinoids on glucose transport and insulin-mediated signaling in skeletal muscles of diabetic (db/db) mice. *J Biol Chem*, **279**, 19721–31.

Shi, Z., Olson, C.A., Bell, A.J. & Kallenbach, N.R. (2001). Stabilization of alpha-helix structure by polar side-chain interactions: complex salt bridges, cation-pi interactions, and C-H-O-H-bonds. *Biopolymers*, **60**, 366–80.

Shi, Z., Olson, C.A., Bell, A.J. & Kallenbach, N.R. (2002). Non-classical helix-stabilizing interactions: C-H-O-H-bonding between Phe and Glu side chains in alpha-helical peptides. *Biophys Chem*, **101-102**, 267–79.

Skulachev, V.P. (1991). Fatty acid circuit as a physiological mechanism of uncoupling of oxidative phosphorylation. *FEBS Lett*, **294**, 158–62.

Slidel, T. (1997). A computational study of chirality in protein structure. *PhD thesis, University of London, London.*

Smith, S., Ee, H., Conners, J. & German, M. (1999). Paired-homeodomain transcription factor PAX4 acts as a transcriptional repressor in early pancreatic development. *Mol Cell Biol*, **19**, 8272–80.

Snow, M. & Larsen, P. (2000). Structure and expression of daf-12: a nuclear hormone receptor with three isoforms that are involved in development and aging in Caenorhabditis elegans. *Biochim Biophys Acta*, **1494**, 104–16.

Solanes, G., Vidal-Puig, A., Grujic, D., Flier, J.S. & Lowell, B.B. (1997). The human uncoupling protein-3 gene. Genomic structure, chromosomal localization, and genetic basis for short and long form transcripts. *J Biol Chem*, **272**, 25433–6.

Son, C., Hosoda, K., Matsuda, J., Fujikura, J., Yonemitsu, S., Iwakura, H., Masuzaki, H., Ogawa, Y., Hayashi, T., Itoh, H., Nishimura, H., Inoue, G., Yoshimasa, Y., Yamori, Y. & Nakao, K. (2001). Up-regulation of uncoupling protein 3 gene expression by fatty acids and agonists for PPARs in L6 myotubes. *Endocrinology*, **142**, 4189–94.

Sonntag, W.E., Lynch, C.D., Cefalu, W.T., Ingram, R.L., Bennett, S.A., Thornton, P.L. & Khan, A.S. (1999). Pleiotropic effects of growth hormone and insulin-like growth factor (IGF)-1 on biological aging: inferences from moderate caloric-restricted animals. *J Gerontol A Biol Sci Med Sci*, **54**, B521–38.

Soulimane, T., Buse, G., Bourenkov, G.P., Bartunik, H.D., Huber, R. & Than, M.E. (2000). Structure and mechanism of the aberrant ba(3)-cytochrome c oxidase from thermus thermophilus. *EMBO J*, **19**, 1766–76.

Staden, R. (1984). Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res*, **12**, 505–19.

Stefan, N. & Stumvoll, M. (2002). Adiponectin–its role in metabolism and beyond. *Horm Metab Res*, **34**, 469–74.

Stevens, T.J. & Arkin, I.T. (1999). Are membrane proteins "inside-out" proteins? *Proteins*, **36**, 135–43.

Stevens, T.J. & Arkin, I.T. (2000). Turning an opinion inside-out: Rees and Eisenberg's commentary (Proteins 2000;38:121-122) on "Are membrane proteins 'inside-out' proteins?" (Proteins 1999;36:135-143). *Proteins*, **40**, 463–4.

Stevens, T.J. & Arkin, I.T. (2001). Substitution rates in alpha-helical transmembrane proteins. *Protein Sci*, **10**, 2507–17.

Stock, D., Leslie, A. & Walker, J. (1999). Molecular architecture of the rotary motor in ATP synthase. *Science*, **286**, 1700–5.

Stroebel, D., Choquet, Y., Popot, J. & Picot, D. (2003). An atypical haem in the cytochrome b(6)f complex. *Nature*, **426**, 413–8.

Stroumbakis, N., Li, Z. & Tolias, P. (1996). A homolog of human transcription factor NF-X1 encoded by the Drosophila shuttle craft gene is required in the embryonic central nervous system. *Mol Cell Biol*, **16**, 192–201.

Stuart, J.A., Harper, J.A., Brindle, K.M., Jekabsons, M.B. & Brand, M.D. (2001a). A mitochondrial uncoupling artifact can be caused by expression of uncoupling protein 1 in yeast. *Biochem J*, **356**, 779–89.

Stuart, J.A., Harper, J.A., Brindle, K.M., Jekabsons, M.B. & Brand, M.D. (2001b). Physiological levels of mammalian uncoupling protein 2 do not uncouple yeast mitochondria. *J Biol Chem*, **276**, 18633–9.

Styren, S., Bowser, R. & Dekosky, S. (1997). Expression of fetal ALZ-50 reactive clone 1 (FAC1) in dentate gyrus following entorhinal cortex lesion. *J Comp Neurol*, **386**, 555–61.

Sui, H., Han, B., Lee, J., Walian, P. & Jap, B. (2001). Structural basis of water-specific transport through the AQP1 water channel. *Nature*, **414**, 872–8.

Sulistijo, E., Jaszewski, T. & MacKenzie, K. (2003). Sequence-specific dimerization of the transmembrane domain of the "BH3-only" protein BNIP3 in membranes and detergent. *J Biol Chem*, **278**, 51950–6.

Sun, J. & Tower, J. (1999). FLP recombinase-mediated induction of Cu/Zn-superoxide dismutase transgene expression can extend the life span of adult Drosophila melanogaster flies. *Mol Cell Biol*, **19**, 216–28.

Tagle, D., Koop, B., Goodman, M., Slightom, J., Hess, D. & Jones, R. (1988). Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol*, **203**, 439–55.

Tatar, M., Kopelman, A., Epstein, D., Tu, M.P., Yin, C.M. & Garofalo, R.S. (2001). A mutant Drosophila insulin receptor homolog that extends life-span and impairs neuroendocrine function. *Science*, **292**, 107–10.

Taylor, W.R., Jones, D.T. & Green, N.M. (1994). A method for alpha-helical integral membrane protein fold prediction. *Proteins*, **18**, 281–94.

Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**, 4673–80.

Thornton, J., MacArthur, M., McDonald, I., Jones, D., Mitchell, J., Nandi, C., Price, S. & Zvelebil, J. (1993). Protein structures and complexes: what they reveal about the interactions which stabilise them. *Phil Trans Roy Soc*, **345**, 113–129.

Tissenbaum, H. & Ruvkun, G. (1998). An insulin-like signaling pathway affects both longevity and reproduction in Caenorhabditis elegans. *Genetics*, **148**, 703–17.

Toyoshima, C. & Nomura, H. (2002). Structural changes in the calcium pump accompanying the dissociation of calcium. *Nature*, **418**, 605–11.

Toyoshima, C., Nakasako, M., Nomura, H. & Ogawa, H. (2000). Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 A resolution. *Nature*, **405**, 647–55.

Trezeguet, V., Le Saux, A., David, C., Gourdet, C., Fiore, C., Dianoux, A., Brandolin, G. & Lauquin, G.J. (2000). A covalent tandem dimer of the mitochondrial ADP/ATP carrier is functional in vivo. *Biochim Biophys Acta*, **1457**, 81–93.

Tsai, J., Taylor, R., Chothia, C. & Gerstein, M. (1999). The packing density in proteins: standard radii and volumes. *J Mol Biol*, **290**, 253–66.

Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R. & Yoshikawa, S. (1996). The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 A. *Science*, **272**, 1136–44.

Tu, N., Chen, H., Winnikes, U., Reinert, I., Marmann, G., Pirke, K.M. & Lentes, K.U. (1999). Structural organization and mutational analysis of the human uncoupling protein-2 (hUCP2) gene. *Life Sci*, **64**, PL41–50.

Ulmschneider, M.B. & Sansom, M.S. (2001). Amino acid distributions in integral membrane protein structures. *Biochim Biophys Acta*, **1512**, 1–14.

Urbankova, E., Hanak, P., Skobisova, E., Ruzicka, M. & Jezek, P. (2003). Substitutional mutations in the uncoupling protein-specific sequences of mitochondrial uncoupling protein UCP1 lead to the reduction of fatty acid-induced H+ uniport. *Int J Biochem Cell Biol*, **35**, 212–20.

Valdar, W.S. & Thornton, J.M. (2001). Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins*, **42**, 108–24.

van de Vossenberg, J.L., Albers, S.V., van der Does, C., Driessen, A.J. & van Klompenburg, W. (1998). The positive inside rule is not determined by the polarity of the delta psi (transmembrane electrical potential). *Mol Microbiol*, **29**, 1125–7.

Van den Berg, B., Clemons, J.r., Collinson, I., Modis, Y., Hartmann, E., Harrison, S. & Rappoport, T. (2004). X-ray structure of a protein-conducting channel. *Nature*, **427**, 36–44.

van Klompenburg, W., Nilsson, I., von Heijne, G. & de Kruijff, B. (1997). Anionic phospholipids are determinants of membrane protein topology. *EMBO J*, **16**, 4261–6.

Vidal-Puig, A.J., Grujic, D., Zhang, C.Y., Hagen, T., Boss, O., Ido, Y., Szczepanik, A., Wade, J., Mootha, V., Cortright, R., Muoio, D.M. & Lowell, B.B. (2000). Energy metabolism in uncoupling protein 3 gene knockout mice. *J Biol Chem*, **275**, 16258–66.

Vijayan, M., Raptis, S. & Sathiyaa, R. (2003). Cortisol treatment affects glucocorticoid receptor and glucocorticoid-responsive genes in the liver of rainbow trout. *Gen Comp Endocrinol*, **132**, 256–63.

Viteri, G., Carrard, G., Birlouez-Aragon, I., Silva, E. & Friguet, B. (2004). Age-dependent protein modifications and declining proteasome activity in the human lens. *Arch Biochem Biophys*, **427**, 197–203.

von Heijne (1991). Proline kinks in transmembrane alpha-helices. *J Mol Biol*, **218**, 499–503.

von Heijne, G. & Gavel, Y. (1988). Topogenic signals in integral membrane proteins. *Eur J Biochem*, **174**, 671–8.

Wadekar, S., Li, D., Periyasamy, S. & Sanchez, E. (2001). Inhibition of heat shock transcription factor by GR. *Mol Endocrinol*, **15**, 1396–410.

Wadekar, S., Li, D. & Sanchez, E. (2004). Agonist-activated glucocorticoid receptor inhibits binding of heat shock factor 1 to the heat shock protein 70 promoter in vivo. *Mol Endocrinol*, **18**, 500–8.

Walder, K., Norman, R.A., Hanson, R.L., Schrauwen, P., Neverova, M., Jenkinson, C.P., Easlick, J., Warden, C.H., Pecqueur, C., Raimbault, S., Ricquier, D., Silver, M.H., Shuldiner, A.R., Solanes, G., Lowell, B.B., Chung, W.K., Leibel, R.L., Pratley, R. & Ravussin, E. (1998). Association between uncoupling protein polymorphisms (UCP2-UCP3) and energy metabolism/obesity in Pima indians. *Hum Mol Genet*, **7**, 1431–5.

Walker, G. & Lithgow, G. (2003). Lifespan extension in C. elegans by a molecular chaperone dependent upon insulin-like signals. *Aging Cell*, **2**, 131–9.

Walker, J.E. & Runswick, M.J. (1993). The mitochondrial transport protein superfamily. *J Bioenerg Biomembr*, **25**, 435–46.

Wallin, E. & von Heijne, G. (1998). Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci*, **7**, 1029–38.

Wang, J. & Kim, S. (2003). Global analysis of dauer gene expression in Caenorhabditis elegans. *Development*, **130**, 1621–34.

Wang, T. & Stormo, G. (2003). Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, **19**, 2369–80.

Waxman, D. (1999). P450 gene induction by structurally diverse xenochemicals: central role of nuclear receptors CAR, PXR, and PPAR. *Arch Biochem Biophys*, **369**, 11–23.

Weihua, X., Lindner, D. & Kalvakolanu, D. (1997). The interferon-inducible murine p48 (ISGF3gamma) gene is regulated by protooncogene c-myc. *Proc Natl Acad Sci (U S A)*, **94**, 7227–32.

Weill, L., Shestakova, E. & Bonnefoy, E. (2003). Transcription factor YY1 binds to the murine beta interferon promoter and regulates its transcriptional capacity with a dual activator/repressor role. *J Virol*, **77**, 2903–14.

Weiss, M. & Schulz, G. (1992). Structure of porin refined at 1.8 A resolution. *J Mol Biol*, **227**, 493–509.

Werck-Reichhart, D. & Feyereisen, R. (2000). Cytochromes P450: a success story. *Genome Biol*, **1**, REVIEWS3003.

Whelan, J., Cordle, S., Henderson, E., Weil, P. & Stein, R. (1990). Identification of a pancreatic beta-cell insulin gene transcription factor that binds to and appears to activate cell-type-specific expression: its possible relationship to other cellular factors that bind to a common insulin gene sequence. *Mol Cell Biol*, **10**, 1564–72.

White, S.H. (2001). Tryptophan and the folding of proteins into membranes. *Biophys J*, **80**, 121–132.

Whittington, D., Waheed, A., Ulmasov, B., Shah, G., Grubb, J., Sly, W. & Christianson, D. (2001). Crystal structure of the dimeric extracellular domain of human carbonic anhydrase XII, a bitopic membrane protein overexpressed in certain cancer tumor cells. *Proc Natl Acad Sci U S A*, **98**, 9545–50.

Wiggins, P. & Phillips, R. (2004). Analytic models for mechanotransduction: gating a mechanosensitive channel. *Proc Natl Acad Sci (U S A)*, **101**, 4071–6.

Williams, G.C. (1957). Pleiotropy, natural selection and the evolution of senscense. *Evolution*, **11**, 398–411.

Williamson, I., Alvis, S., East, J. & Lee, A. (2003). The potassium channel KcsA and its interaction with the lipid bilayer. *Cell Mol Life Sci*, **60**, 1581–90.

Wilson, D., Johnson, P. & McCord, B. (2001). Nonradiochemical DNase I footprinting by capillary electrophoresis. *Electrophoresis*, **22**, 1979–86.

Wimley, W.C., Creamer, T.P. & White, S.H. (1996). Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides. *Biochemistry*, **35**, 5109–24.

Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhauser, R., Pruss, M., Schacherer, F., Thiele, S. & Urbach, S. (2001). The TRANSFAC system on gene expression regulation. *Nucleic Acids Res*, **29**, 281–3.

Winkler, E., Wachter, E. & Klingenberg, M. (1997). Identification of the pH sensor for nucleotide binding in the uncoupling protein from brown adipose tissue. *Biochemistry*, **36**, 148–55.

Wolkow, C.A., Kimura, K.D., Lee, M.S. & Ruvkun, G. (2000). Regulation of C. elegans life-span by insulinlike signaling in the nervous system. *Science*, **290**, 147–50.

Woolfson, D.N. & Williams, D.H. (1990). The influence of proline residues on alpha-helical structure. *FEBS Lett*, **277**, 185–8.

Woolfson, D.N., Mortishire-Smith, R.J. & Williams, D.H. (1991). Conserved positioning of proline residues in membrane-spanning helices of ion-channel proteins. *Biochem Biophys Res Commun*, **175**, 733–7.

Wotton, D., Lo, R., Lee, S. & Massague, J. (1999a). A Smad transcriptional corepressor. *Cell*, **97**, 29–39.

Wotton, D., Lo, R., Swaby, L. & Massague, J. (1999b). Multiple modes of repression by the Smad transcriptional corepressor TGIF. *J Biol Chem*, **274**, 37105–10.

Wozniak, A., Drewa, G., Wozniak, B. & Schachtschabel, D. (2004). Activity of antioxidant enzymes and concentration of lipid peroxidation products in selected tissues of mice of different ages, both healthy and melanoma-bearing. *Z Gerontol Geriatr*, **37**, 184–9.

Xia, D., Yu, C., Kim, H., Xia, J., Kachurin, A., Zhang, L., Yu, L. & Deisenhofer, J. (1997). Crystal structure of the cytochrome bc1 complex from bovine heart mitochondria. *Science*, **277**, 60–6.

Xiao, N. & DeFranco, D. (1997). Overexpression of unliganded steroid receptors activates endogenous heat shock factor. *Mol Endocrinol*, **11**, 1365–74.

Yankovskaya, V., Horsefield, R., Tornroth, S., Luna-Chavez, C., Miyoshi, H., Leger, C., Byrne, B., Cecchini, G. & Iwata, S. (2003). Architecture of succinate dehydrogenase and reactive oxygen species generation. *Science*, **299**, 700–4.

Yeates, T., Komiya, H., Rees, D., Allen, J. & Feher, G. (1987). Structure of the reaction center from Rhodobacter sphaeroides R-26: membrane-protein interactions. *Proc Natl Acad Sci U S A*, **84**, 6438–42.

Yu, E., McDermott, G., Zgurskaya, H., Nikaido, H. & Koshland, J.r. (2003). Structural basis of multiple drug-binding capacity of the AcrB multidrug efflux pump. *Science*, **300**, 976–80.

Yuan, J., Tirabassi, R., Bush, A. & Cole, M. (1998). The C. elegans MDL-1 and MXL-1 proteins can functionally substitute for vertebrate MAD and MAX. *Oncogene*, **17**, 1109–18.

Yuen, C.T., Davidson, A.R. & Deber, C.M. (2000). Role of aromatic residues at the lipid-water interface in micelle-bound bacteriophage M13 major coat protein. *Biochemistry*, **39**, 16155–62.

Yuh, C., Bolouri, H. & Davidson, E. (1998). Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, **279**, 1896–902.

Zamiri, M. (1978). Effects of reduced food intake on reproduction in mice. *Aust J Biol Sci*, **31**, 629–39.

Zhang, Z., Huang, L., Shulmeister, V., Chi, Y., Kim, K., Hung, L., Crofts, A., Berry, E. & Kim, S. (1998). Electron transfer by domain movement in cytochrome bc1. *Nature*, **392**, 677–84.

Zhou, Y., Morais-Cabral, J., Kaufman, A. & MacKinnon, R. (2001). Chemistry of ion coordination and hydration revealed by a K+ channel-Fab complex at 2.0 A resolution. *Nature*, **414**, 43–8.

Zouni, A., Witt, H., Kern, J., Fromme, P., Krauss, N., Saenger, W. & Orth, P. (2001). Crystal structure of photosystem II from Synechococcus elongatus at 3.8 A resolution. *Nature*, **409**, 739–43.

Zwaan, B., Bijlsma, R. & Hoekstra, R.F. (1995). Direct selection on life span in Drosophila melanogaster. *Evolution*, **49**, 649–659.